BEHAVIORAL AND BRAIN SCIENCES (1999) **22,** 127–196
Printed in the United States of America

A connectionist theory of phenomenal experience

Gerard O'Brien
Jonathan Opie
Department of Philosophy, The University of Adelaide, Adelaide, South Australia 5005, Australia.
gobrien@arts.adelaide.edu.au
chomsky.arts.adelaide.edu.au/Philosophy/gobrien.htm
jopie@arts.adelaide.edu.au
chomsky.arts.adelaide.edu.au/Philosophy/jopie.htm

Abstract: When cognitive scientists apply computational theory to the problem of phenomenal consciousness, as many have been doing recently, there are two fundamentally distinct approaches available. Consciousness is to be explained either in terms of the nature of the representational vehicles the brain deploys or 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, although 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 that 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. 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 used in the dissociation studies – so critical, in fact, that it is 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 target article we develop and defend this connectionist vehicle theory of consciousness. It takes the form of the following simple empirical hypothesis: *phenomenal experience consists of the explicit representation of information in neurally realized parallel distributed processing (PDP) networks.* This hypothesis leads us to reassess some common wisdom about consciousness, but, we argue, in fruitful and ultimately plausible ways.

Keywords: classicism; computation; connectionism; consciousness; dissociation; mental representation; phenomenal experience

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, because the vast majority of human beings have such experiences. Furthermore, there is probably something it is like to be a dog, and perhaps even fish have phenomenal experiences, however minimal and fleeting they 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*



Gerard O'Brien completed his D.Phil. degree at the University of Oxford before taking a position in the Department of Philosophy at the University of Adelaide, Australia, where currently he is a senior lecturer and the director of the Graduate Program in Cognitive Science. He has published widely in the fields of philosophy and mind and the foundations of cognitive science, with a particular focus on exploring the connectionist alternative to the classical computational conception of human cognition.



Jonathan Opie recently commenced an Australian postdoctoral research fellowship in the Department of Philosophy at the University of Adelaide, after completing his Ph.D. there. He has published articles on computational approaches to consciousness in a number of major journals.

© 1999 Cambridge University Press 0140-525X/99 $12.50

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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* 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.

However, 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. 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 (Searle [1983, pp. 1–2], e.g., cites pains and undirected emotional experiences in this regard); 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.

The advocate of a such an account of consciousness is not completely without resources here, however. With regard to the first difficulty, for example, 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; 1997] 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 analyze these in terms of the information they carry about occurrences at certain bodily locations (see, e.g., Tye 1996). As for the so-called undirected emotions, it is plausible to analyze 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 centers where the emotion is "felt" (see, e.g., Charland 1995; Johnson-Laird 1988, pp. 372–76; Schwartz 1990).

Moreover, with regard to the second difficulty, although 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 *nonexplicit* forms of information coding. Information encoded in a computational device in such a way that each distinct item of data is encoded by a physically discrete object is typically said to be represented explicitly. Information that is stored in a dispositional fashion, or embodied in a device's primitive computational operations, on the other hand, is said to be represented nonexplicitly.[3] It is reasonable to conjecture that the brain uses 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 us 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. Baars's "Global Workspace" model of consciousness (1988) is a representative example. Baars's 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) because 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, as 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 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 the other way around: 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 choose to explore the former path?

The answer, we suggest, is twofold. First, there is the influence exerted by a large body of research purporting 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 using paradigms such as dichotic listening, visual masking, and implicit learning, as well as the investigation of neurological disorders such as blindsight. Such "dissociation studies,"

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

as we will call them, appear to rule out a vehicle theory. 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 does not 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.

Recent developments in cognitive science combine, however, to suggest that a reappraisal of this situation is in order. On the one hand, a number of theorists have been highly critical of the experimental methodologies used in the dissociation studies. So critical, in fact, that it is no longer reasonable to assume that the dissociability of conscious experience and explicit representation has been adequately demonstrated (see, e.g., Campion et al. 1983; Dulany 1991; Holender 1986; Shanks & St. John 1994; see also Velmans: "Is Human Information Processing Conscious?" *BBS* 14(4) 1991). On the other hand, as a theory of human cognition, classicism 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 that 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 target article is to develop and defend this connectionist vehicle theory of consciousness. We begin, in section 2, with a rapid reevaluation of the dissociation studies. It is not our goal here to provide a thoroughgoing 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 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 target article. 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. The same does not apply to connectionists, however. In the final substantive section of the article (sect. 5), we present and defend a connectionist vehicle theory of consciousness. This theory leads us to reassess some common wisdom about consciousness, but, we argue, in fruitful and ultimately plausible ways.

2. The dissociation studies: A reappraisal

The literature in cognitive science is full of experimental work that 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. The latter, 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," because they examine differential predictions concerning the existence and role of unconscious information in various kinds of thought (1991, p. 107). 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, test 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. 34–35). 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 will not 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). Postexperimental 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 toward the interpretation suggested by the disambiguating context (reported in Holender 1986).

The moral here is fairly obvious. To bias a subject's paraphrase of attended material, the unattended input must clearly undergo processing all the way to the semantic level. 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. However, both the Lackner and Garrett and the MacKay

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

studies suggest that this representation does not evoke any conscious experience. There is prima facie evidence, therefore, for the dissociation of explicit representation and conscious experience.

Not all cognitive psychologists accept the conclusions typically drawn from dichotic listening studies, however (see, e.g., Holender 1986). Indeed, there is reason to believe that the apparent support for the dissociation of explicit representation and phenomenal experience generated by this research is an artefact of poor methodology. For example, there is the reliance on postexperiment 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 by verbal recall, whereas 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). Although 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 striking, 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 caused by uncontrolled attention shifts to the secondary channel, shifts that resulted in brief conscious awareness of the disambiguating context, even if this experience could not 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 1970s. (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 (1986) 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., Perenin 1978; Perenin & Jeannerod 1975, 1978; Weiskrantz 1980; Weiskrantz et al. 1974). There is also evidence that such subjects can discriminate patterns of various kinds. A forced-choice technique has been used, 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 et al. 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 cannot 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. It is reasonable to suppose that visual judgments are mediated by mental representations: 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 the dissociation of explicit representation and conscious experience.

One should not be too hasty however; blindsight research is not without controversy. Campion et al. (1983) argue that none of the existing blindsight studies provides adequate controls for light scatter. Furthermore, they claim that it is 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 have not 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 already raised, 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 reason to believe, therefore, 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, because blindsight subjects putatively have no visual experience of the objects

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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. 435–36). The disagreement here is probably partly caused by equivocation over the use of terms like "aware" and "conscious" (among the researchers), in conjunction with a failure to ask precise 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. In fact, however, the literature contains a great many reports of experiences that co-occur with discriminative episodes. Consider the comments made by Weiskrantz's subject D.B., after performing well above chance on a test that involved distinguishing between Xs and Os presented in his scotoma. Although 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. He talks of things "popping out a couple of inches" and of "moving waves," in response to single-point stimuli (Weiskrantz 1986, p. 45). He also refers to "kinds of pulsation" and of "feeling some movement" in response to moving line stimuli (Weiskrantz 1986, p. 67).

Consequently, although blindsight subjects clearly do not have normal visual experience in the "blind" regions of their visual fields, this is not to say that they do not have any phenomenal experience whatsoever associated with stimuli presented in these regions. Further, 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 he does not need to in order to perform this task. All he requires is some way of discriminating between the two stimulus conditions – some broad phenomenal criterion to distinguish "X-ness" from "O-ness." And as we have seen, he does possess such a criterion: one stimulus condition feels "jagged" whereas the other feels "smooth." It is therefore 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 1997 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). Several 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 already discussed 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] Therefore, it is typically not the stimuli that subjects are held to be unaware of in implicit learning situations. It is, rather, the relationships between the stimuli that are thought to be unconscious (Shanks & St. John 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 another 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 guides subsequent decision-making, even though it remains unconscious.

It may be that the standard interpretation is somewhat incautious, however. Shanks and St. John (1994), in their wide-ranging critique, have identified two principal criteria that implicit learning studies must satisfy to establish unconscious learning (in the sense already specified). First, tests of awareness must be sensitive to all relevant conscious knowledge (the sensitivity criterion); and second, it must be

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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 will not consider the sensitivity criterion in detail here, but note that a great many studies of implicit learning have relied entirely on postexperiment 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. 374–75, for discussion). At any rate, it is the information criterion that 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 et al. (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 1997, p. 193). Similar results have been reported by Perruchet and Pacteau (1990), and Dienes et al. (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 judgments "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 1997, pp. 193–94). Based on their extensive analysis of this literature, Shanks and St. John conclude: "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. 388–89). 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 used to investigate subliminal perception: perceptual integrations that, because of 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 judgments for stimuli between 5 and 10 msec below the supraliminal threshold. Subjects later reported that they sometimes "felt silly" making a judgment about a stimulus they had not 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 judgments, 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 et al. (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 is quite possible (in this stimulus-energy domain) that fluctuations in the visual system will occasionally generate conscious events. Therefore, the (small) degree of priming that occurs may well be entirely a result of 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 have described are methodologi-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

cally 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, unforeseen 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, because the available evidence does not appear to rule out this approach conclusively.

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, fortunately, one 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 refrigerator, 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, 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. This is a relatively weak claim, however. Indeed, given the now standard interpretation of (what has come to be known as) the Church–Turing thesis (see, e.g., Kleene 1967, p. 232) – namely, 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 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. 13–14)

In what follows, we will adopt this strong interpretation of classicism (see also Fodor 1975; 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 section 1, 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, which we believe respects the implicit commitments of most theorists in this area (see also Cummins 1986; Pylyshyn 1984). We will use 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. (Dennett 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. They are thus "explicit representations," according to Dennett's taxonomy, physically distinct objects, each possessed of a single semantic value. In the classical context, explicit representation consists of 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 on these objects that explicit information[11] supervenes, according to the classicist.

Dennett identifies three further styles of representation, which we will refer to collectively as *nonexplicit.* 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" (Dennett 1982, p. 216). It is questionable, however, whether the concept of implicit representation, de-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

fined in this way, is relevant to classical cognitive science. Logical consequences do not have effects unless there are mechanisms whereby a system can derive (and use) them. Also, it is clear from the way Dennett defines it that implicit information can exist in the absence of such mechanisms. Another way of putting this is to say that although 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 that need not be physically instantiated. Therefore, 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 that a computational system is capable of rendering explicit, as *potentially explicit* (1982, pp. 216–17). [See also Searle: "Consciousness, Explanatory Inversion and Cognitive Science," *BBS* 13(4) 1990.] 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, although not yet explicit, is not merely implicit; it is potentially explicit, by 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; Fodor 1987, Ch. 1). In other words, in 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 by 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. 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. (Dennett 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 is of course the basis of stored program digital computers and the possibility of a Universal Turing machine (one that can emulate the behavior of any other Turing machine). However, 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 nonexplicit 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 (because 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 describe 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 insofar as some aspects of brain maturation require the presence of environmental "triggers." Thus, 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 merely a 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: (1) all phenomenal experience is the result

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

of the tokening of symbols in the brain's representational media, and that whenever such symbols are tokened, their contents are phenomenally experienced; and (2) whenever information is causally implicated in cognition, yet not consciously experienced, such information is encoded nonexplicitly.

However, on the face of it, a classicist cannot 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. Recall that in such a system, 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). Although a Turing machine may have this capacity, until it actually renders a piece of information explicit, this information cannot influence the ongoing behavior of the system. In fact, qua potentially explicit, such information is just as causally impotent as the logical entailments of explicit information. For potentially explicit information to be causally efficacious, it must first be physically embodied as symbols written on the machine's tape. 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 is 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" (Pylyshyn 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 will 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] Thus, 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: There is a complex economy of unconscious states that mediate the sequence of conscious episodes. Although it is possible that all the rules of inference are tacit, this mediating train of unconscious beliefs must interact to produce their effects; otherwise we do not have a causal explanation. However, 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 work, p. 25). Therefore, 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. Although 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 by means of the generation and confirmation of hypotheses. This process must be largely unconscious because much of our learning appears not to involve conscious hypothesis testing. As with cognition more generally, it requires an interacting system of unconscious representations, which, for a classicist, 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 learning is always and everywhere merely a process that reconfigures the brain's functional architecture. However, any classicist who claims that learning is noncognitive 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" (Fodor 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. However, 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 tokened in 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 does not 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

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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: 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* (PDP).[16]

A PDP network consists of 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. According to connectionism, human cognitive processes are the 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 that must be addressed, the first concerning the manner in which connectionism differs from classicism, and 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 because 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 do not 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 just described loses its force: although many classicists claim that PDP represents a plausible implementation-level (i.e., noncognitive) framework for classical models of cognition (see, e.g., Fodor & Pylyshyn 1988, pp. 64–66), 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 used 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 & McClelland 1986, Ch. 4). Sejnowski goes even further, arguing that although 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" (Sejnowski 1986, p. 388). Smolensky (1988), on the other hand, argues that because we are still largely ignorant about the dynamic 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.

Our own interpretation of the relationship between PDP systems and real neural networks puts us at the former end of this spectrum (see also Bechtel 1988b; 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 networks 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, Ch. 3). In this respect, the PDP framework isolates and hence enables us to focus on the computationally significant properties of neural networks, while ignoring their fine-grained neurochemistry. Our best neuroscience informs us that neural networks 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 is 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,

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Ch. 3). This does not impugn the integrity of the framework as a whole, however. Moreover, it is entirely open to connectionists to incorporate more complex dynamic features of neural networks, 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 & 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 do not 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. However, 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 used 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 suggests about the way information is encoded in the brain. Although it was formulated in the context of digital computational theory, Dennett's (1982) 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, because each of these patterns is a physically discrete, structurally complex object, which, like the symbols in conventional computers, possesses 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" by 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. Although 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]

Although 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. Therefore, a network, by 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. Such long-term storage of information is superpositional in nature, because 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? 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 of 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, because they completely determine the system's response to input.

4.2 Connectionism and consciousness. With these PDP styles of representation before us, let us now address the key question: Does connectionism have the computational resources to support a vehicle theory of consciousness? As was the case with classicism, 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 nonexplicitly, in the form of potentially explicit/tacit representations.

Is this suggestion any more plausible in its connectionist incarnation than in the classical context? We think it is. In the next section we will 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.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Although we were able to apply Dennett's (1982) 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 nonexplicit representational resources, is unable to meet all the causal demands cognition places on the unconscious (and is therefore 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 by 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 are 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. There is a strong sense, therefore, 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 because, 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 1993a.) 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 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. Furthermore, learning, in the connectionist theory, involves the progressive modification of a network's connection weights and pattern of connectivity, to encode further potentially explicit information. Learning, in other words, is a process that 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 nonexplicit 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 do not 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 nonexplicit representation. Explicit representations, in this alternative conception, are the products of unconscious processes, and therefore a connectionist can feel encouraged by 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 network "relaxation" processes that are determined by the superpositionally encoded information stored therein, and result in stable patterns of activation. Unconscious processes thus generate activation pattern representations, which the connectionist is free to identify with individual phenomenal experiences, because 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.

Thus, connectionism does appear to have the right computational profile to permit a vehicle theory of consciousness. Because such theories are all but absent from contemporary cognitive science, we believe it is worth exploring this much neglected region of the theoretical landscape. In the next section, we do just that 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 counterintuitive 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 nonexplicit representation to account for unconscious thought processes and learning, it is possible to align phenomenal experience with explicit information coding in the brain. That, baldly stated, is the connectionist vehicle theory of consciousness we want 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

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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 (Chalmers 1995; 1996; Nagel 1974).

5.1. The intranetwork level. The connectionist account of consciousness we have proposed is not completely novel. Theorists involved in laying the foundation 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 (McClelland & Rumelhart 1986; Rumelhart & McClelland 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 et al. 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" (Smolensky 1988a, p. 13).

It is worth pointing out, however, that neither Rumelhart et al. (1986) nor Smolensky (1988) takes the presence of a stable pattern of activation to be both necessary and sufficient for consciousness. Rumelhart et al. do not appear to regard stability as necessary for consciousness, because 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" (Rumelhart et al. 1986, p. 39). This implies, however, that "on occasions in which the relaxation process is especially slow, consciousness will be the time average over a dynamically changing set of patterns" (Rumelhart et al. 1986, p. 39). In other words, stability is not necessary for conscious experience, because even a network that has not yet stabilized will, on this account, give rise to some form of consciousness. Smolensky, on the other hand, does not regard stable activation to be sufficient for consciousness, and says as much (Smolensky 1988a, 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 1997). Furthermore, 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 et al. (1986) can be used to accommodate vague, fleeting, and peripheral forms of experience (what, following James [1890], he calls the "fringe" of consciousness) within a computational framework (see Mangan 1993b). Like Rumelhart et al. (1986), however, Mangan seems to accept the possibility that states of consciousness could be associated with networks that have not fully stabilized – that is, with stabilizing networks – rather than restricting them to stable patterns of activation across such networks.

Finally, Lloyd (1991; 1995; 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" (Lloyd 1995a, p. 165). Lloyd provides a detailed analysis of phenomenal experience, developing the distinctions between sensory and nonsensory, 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 cashed out in connectionist terms (1995; 1996). Again, Lloyd appears to focus his efforts on activation patterns in general, rather than on 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, although 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 is possible for real neural networks to generate many stable states per second (Churchland & Sejnowski 1992, Ch. 2). Consequently, what is a rapid sequence of stable patterns at the level of an individual neural network may be a continuous phenomenal stream at the level of consciousness.

The second point 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. This data structure is not

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

equivalent to a pattern of activation across a real (nonsimulated) PDP network, however. 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 used 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 with 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 activation patterns are absent in digital simulations of PDP systems, so are phenomenal experiences, on our account.

There are further reasons to focus on stable activation patterns, 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 does not 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. However, stable input is the result of stable output. Thus, one network can contribute to the generation of explicit information in another only if it is 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 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, intranetwork property of the brain's neural networks. On this account, therefore, internetwork information processing relations depend on phenomenal experience, not the reverse.[24] Moreover, the presence of phenomenal experience is necessary, but not sufficient, for such internetwork 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, although 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 internetwork information processing. This is a much neglected region in connectionist theorizing, most of which tends to focus on intranetwork activity.[25] In the next subsection we will 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 internetwork 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. This is surely a mistake. Many lines of evidence suggest that there is a significant degree of modularity in brain architecture. Connectionism is constrained by this 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. However, because 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, because the neural vehicles of explicit representation appear to be very numerous, the connectionist vehicle theory of consciousness also implies that the neurological basis of consciousness is manifold, that is, that there are a multitude of consciousness-making mechanisms in the brain.

Although there are those who might be prepared to reject one or the other of these implications, we suggest that they are quite consistent with the existing evidence, both phenomenological and neurological. Consider first the ev-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

idence 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 printed sentences before you, your (global) phenomenal experience is simultaneously multimodal and multichannelled: 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 do not momentarily disappear: you do not stop feeling where your limbs are; you do not stop having auditory experiences; you do not 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 is 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. However, people must often suffer severe neurological damage before they can even acknowledge the existence of other persisting aspects of this field. For example, Sacks describes the tragic case of a woman who, because of 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" (Sacks 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 on 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 that 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, because they run side by side, the loss of any one strand does not 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 modalities. Consider the familiar "inverting stairs" ambiguous figure (Fig. 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 that is subject to a degree of voluntary control. Or consider the "vase/faces" ambiguous figure (Fig. 2). Whether one interprets it as a vase (light figure, dark background), or as a pair of faces (dark figure, light 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 a further 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.

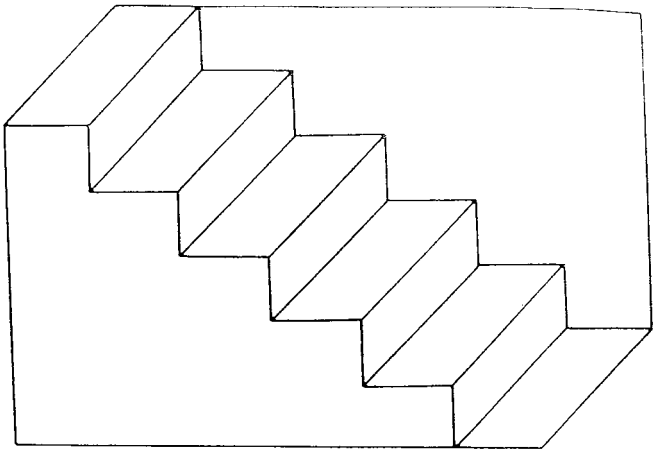


Figure 1. Inverting stairs ambiguous figure.

The real force of this phenomenological evidence emerges fully only when it is conjoined with the available neuroscientific evidence. We know, on the basis of deficit studies, that the information processing that supports con-

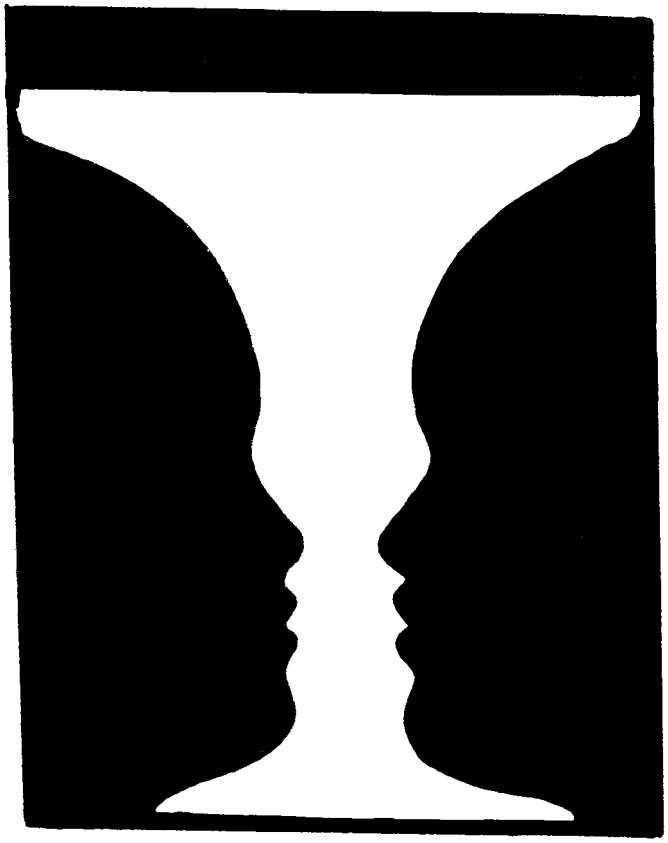


Figure 2. Vase/faces ambiguous figure.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

scious experience is realized in structures distributed right across the brain, and that the distributed nature of this information processing is both an intramodal and an intermodal 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, because of a vascular disorder in the brain that 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. (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. Equally striking case studies are available in relation to the loss of color sensations (see, e.g., 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 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]

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. According to 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. Also according to 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 nonsensory kinds of experience (Lloyd 1996). According to Lloyd, sensory experiences, unlike nonsensory 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. 65–67). 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 nonbasic. 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 herein. 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) that 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 lose something – a quality of extension let us say – that 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, but 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 already described because we are describing an element of experience beyond the mere perception of a face as an organized whole. A familiar face is perceived not only 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 before 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."

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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.) Although 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; 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 is 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, and feelings of facial familiarity, to highly abstract language-based understanding experiences. We suggest that there is a natural structural feature of the brain that the connectionist vehicle theory of consciousness can use 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, whereas others are second-order (i.e., they receive their input from the first layer of networks), and so on. It is natural to suppose, according to the connectionist vehicle theory of consciousness, that less abstract elements of experience correspond to stable patterns of activation in lower-order networks, whereas 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 corresponds 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.

5.3. The unity of consciousness. Despite the compelling support for the connectionist vehicle theory that we have 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. 398–99); and 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.

However, if consciousness is just one thing, then there must be one thing that underlies it. Because 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 (1995), 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 "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 to become conscious (the thalamic projection system and associated structures). These brain structures realize an executive system that 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. What underlies consciousness, therefore, is not one thing, but many. We might refer to this as a *multitrack* 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 consciousness-making system. On the face of it, a multitrack model renders the unity of consciousness somewhat mysterious. A single-track model, on the other hand, is in the business of rendering consciousness unitary.

It is pertinent at this point, however, 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 dependent 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. If this is the sense in which consciousness is unified, then it is 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. To do so, it will first be necessary to delve into the notion of coherence a little further.

Phenomenal experience exhibits both intramodal and in-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

termodal coherence. In our daily experience we sometimes have only one source of information regarding external objects: we hear the bird, but we cannot see it; we see the ball (on the roof), but we cannot feel it. In these cases, we do not 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 by means of 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 these events; the location of these same keystrikes, as revealed in visual experience, is compatible with their position in "auditory space"; and one's proprioceptive and visual experiences of hand position are consonant. Intermodal 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 covary in a consistent fashion.

It is important to recognize that there are two aspects to coherence: the first is temporal coherence, as exemplified in the coincidence of visual, auditory, and tactile experiences of typing. The second is spatial coherence, which manifests itself in numerous ways: 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 multitrack 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 a consequence of the simultaneity of their vehicles (this suggestion is not new; see, e.g., Edelman 1989). So when a felt keystroke is temporally aligned with its seen counterpart in experience, we simply propose to explain this in terms of a brain architecture that generates simultaneous vehicles in those two modalities. It is 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, because when we seek to explain this form of coherence we must contend with what Akins calls the "spatial binding problem," namely: "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?" (Akins 1996, p. 30). Single-track theories of consciousness take this problem in their stride by refusing to identify visual experience solely with the machinations 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. However, a multitrack theorist is in the business of identifying experience with the neural vehicles of explicit information, so the binding problem is pressing. It is not clear, however, that this problem is intractable from the perspective of the connectionist vehicle theory of consciousness. Indeed, it may be no more than a pseudoproblem 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" (Akins 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). This surely is a naive theory. We do not 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 intermodally and intramodally coherent. In particular, the only intramodal "binding" we can reasonably expect is a binding at the level of contents. 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, for example, that each element of visual experience, in addition to its nonspatial content, also incorporates spatial information. That is, the basic elements of vision are *color-x-at-location-y,* and so on.) 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 have not yet touched on another 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 harboring "the contents of the several basic sensory modalities within a *single unified experience"* (Churchland 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 multitrack nature of our account of consciousness, we must address the issue of 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 centers 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 feelings 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. Although 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."

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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. Rejection of the Naive Theory of 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. However, some will find this suggestion objectionable for the reason that it does not seem to provide a satisfying reductive explanation of consciousness. A reductive explanation is satisfying when there is a "perspicuous nexus" between the postulated micromechanism and the macrophenomenon 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 constitute the familiar 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 (Chalmers 1995; 1996; Nagel 1974), 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 is no worse off, in this respect, than any other current theory, but there is more we can say. What we can properly conceive is not fixed, but narrows with the development of our scientific understanding. What today is imaginable might tomorrow merely indicate that we possessed insufficient information. To borrow an example of Cottrell's, anyone 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. (Cottrell 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, note 5 – see also Flanagan 1992, p. 59; 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 eventually be inconceivable that a creature with this mechanism would not be conscious. 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 toward 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. What might not be so readily apparent, however, 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" that 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, Ch. 4; Clark 1993a; Rumelhart & McClelland 1986, Ch. 1–3). The spiking activity across a neural network can be represented in terms of a hyperdimensional activation space, the points of which describe individual activation patterns. 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 to stable patterns of activation that map onto widely separated points in this activation space, whereas points that are near neighbors in this space correspond to 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 one neural substrate is capable of generating all the different kinds of experience we are capable of entertaining (both within and across sensory modalities). 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. 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 distinguish between the various kinds of linguistically mediated experiences in a systematic fashion.

All these differences are explicable, we think, with the resources of connectionist activation pattern representations.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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 the relevant neural network, and "shape," which 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). Equally, however, what distinguishes them from experiences of, say, sound, are these same geometric properties, properties that distinguish one neural network from another, and hence, in our account, one kind of phenomenology from another.

Naturally, 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.[32] However, precisely because this connectionist vehicle theory has 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 target article, we have done something that is singularly unpopular in contemporary cognitive science: we have 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 by 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 nonexplicitly 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 reassess 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. 454–55 makes a similar point). However, 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 across the brain, is precisely the sort of account we need.

Beyond these incentives, we believe 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 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 BBS referees for their very helpful comments on earlier versions of this article. We are also grateful to many audiences at talks on this material for their criticisms.

NOTES

1. In speaking of "phenomenal experiences," our intended target is neither self-consciousness nor what has come to be called access-consciousness (Block 1993; 1995). It is, rather, phenomenal consciousness: the "what it is like" of experience (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 & Schwarz 1991, p. 64; Dietrich 1989; Fodor 1975, p. 27; Von Eckardt 1993, pp. 97–116).

3. See, for example, Cummins 1986; Dennett 1982; Pylyshyn 1984. We discuss the distinction between explicit and nonexplicit representation more fully in section 3.

4. See, for example, Baars 1988; Churchland 1995; Crick 1984; Dennett 1991; Flanagan 1992; Jackendoff 1987; Johnson-Laird 1988; Kinsbourne 1988; 1995; Mandler 1985; Newman 1995; Rey 1992; Schacter 1989; Shallice 1988a; 1988b; and Umiltà 1988.

5. Strictly speaking, there is a third alternative, one that combines these two strategies. On this view, consciousness is to be explained in terms of 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 although process theories have been much debated in cognitive science, vehicle theories have not yet been investigated in any real depth. We aim, in this article, to raise the profile of this alternative strategy.

6. We are assuming here that connectionism does constitute a

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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; Von Eckardt 1993, Ch. 3).

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

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

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 Chs. 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 will not buy into these here (although see Cleland 1993, Copeland 1997, and Rubel 1989 for interesting discussions).

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.

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. Strictly speaking, however, 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.

13. Pylyshyn's (1984) notion of the brain's "functional architecture" arguably incorporates tacit representation. Both he and Fodor (see, e.g., 1987, Ch. 1) 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 is clear that some processing rules must be tacit, or the system could not operate.

14. Alhough 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 (AI), 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 access, update, and process that information very rapidly. This becomes particularly acute for AI when it manifests itself as the "frame problem." See Dennett (1984) for an illuminating discussion.

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

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; 1996), Rumelhart and McClelland (Rumelhart & McClelland 1986; McClelland & Rumelhart 1986), Smolensky (1988), and the earlier Van Gelder (1990). For useful introductions to connectionism, see Bechtel and Abrahamsen 1991; Clark 1989, Chs. 5–6; Rumelhart 1989; Tienson 1987.

18. We say that this is the main argument for this deflationary interpretation of connectionism, but it is hard to find any explicit formulation in published work, although one certainly comes across it in e-mail 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, as well. The only real difference is that in the former case, but not the latter, one computational device is being used 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 sect. 5.1).

21. For good general introductions to the representational properties of PDP systems, see Bechtel and Abrahamsen 1991, Ch. 2; Churchland 1995; Churchland and Sejnowski 1992, Ch. 4; Rumelhart 1989; and Rumelhart and McClelland 1986, Chs. 1-3. More fine-grained discussions of the same can be found in Clark 1993a, and Ramsey et al. 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, is 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; Van Gelder 1990).

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.

24. This intimate relationship between internetwork 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; Dennett 1991). Because 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. However, 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. Some important exceptions here are Clark and Karmiloff-Smith 1993 and Clark and Thornton 1997.

26. Some will object to these claims on the grounds that consciousness is coextensive 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 does not 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

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

information already extracted from the world, and hence already displayed in the phenomenal field, to more intense processing. Such additional processing would engage 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 (Jackendoff 1987, pp. 280–83).

27. There are echoes here of Dennett's multiple drafts theory of consciousness (Dennett 1991; 1993). Like us, Dennett 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" (Dennett 1991, pp. 253–54). 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 neuropsychological evidence pointing to the distributed neural basis of consciousness, see, e.g., the papers in Milner & Rugg [1992].)

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. Individuals with prosopagnosia 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.

29. Theorists who assert this do not necessarily take instantaneous consciousness to be restricted to a single modality. Churchland, for example, regards our "single unified experience" as polymodal in character (Churchland 1995, pp. 214–22).

30. This account is strikingly similar to Baars's "Global Workspace" model of consciousness that we described earlier (see sect. 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.

31. What we're suggesting here is that for us to be able to respond appropriately to rapidly changing local conditions, the various determinants of a behavioral response (visual input, tactile input, proprioceptive input, and so forth) will need to be brought to bear roughly synchronously, so that they do 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.)

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

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Consciousness and agency: Explaining what and explaining who

Richard A. Carlson

Department of Psychology, Penn State University, University Park, PA 16802.
cvy@psu.edu gandalf.la.psu.edu/rich/

Abstract: The target article offers an intriguing hypothesis relating the content of phenomenal experience to a qualitative characteristic of information processing. This hypothesis, however, offers only an explanation of the "what" of consciousness, not the "who" – the experiencing agent remains mysterious. Their hypothesis about the unity of consciousness can be linked to an informational account of the agency or subjectivity of consciousness.

The authors of the target article offer an intriguing hypothesis relating the content of phenomenal experience to a qualitative property of information processing in neural networks, and they argue convincingly that classical computationalism lacks the theoretical resources to provide an analogous account. I think, however, that this hypothesis does not amount to a theory of consciousness, because explaining why or how some subset of the information processed in the nervous system can be identified with contents of consciousness addresses only part of the puzzle. The "what" of phenomenal experience is certainly an important topic, and it is of course essential that any system proposed to underlie phenomenal experience be capable of representational richness commensurate with the richness of phenomenal experience (sect. 5.4). At least as important, though, is the "who" of consciousness – a theory of consciousness ought to explain the informational or computational basis of subjectivity and conscious agency. O'Brien & Opie's (O&O's) account of consciousness represents a significant advance over the more common classicist accounts, but like those accounts it generally adopts what I have called an *implicit agent* stance – the subjectivity or point of view from which conscious contents are considered is left implicit and unexplained by the theory (Carlson 1997). Phenomenal experience comprises contents considered from some point of view. But only near the end of the article (sect. 5.3) do the authors touch briefly on how subjectivity might be conceived in their theory; throughout most of the article, the focus is on contents and the point of view is left implicit. For example, it is mysterious to me why, on the basis of this theory, "networks that receive input from many sources, and are thus least modality-specific" (sect. 5.2, last para.) are *therefore* most subject to voluntary control. Agency is assumed, but not made explicit by the theory.

In my view, a theory of consciousness should explain how a computational framework can account for the existence and activity of conscious agents (Carlson 1992; 1997). I think that such a theory depends on an analysis of information processing that, like Gibson's (1979) theory of visual perception, identifies the self as informationally specified in the course of perceptually and cognitively guided action (Neisser 1988). Furthermore, a theory of consciousness should address perceptual-motor, symbolic, and emotional aspects of awareness, and should offer an account of the conscious control of activity. One aspect of such an account is a description of the processing function of consciousness: how consciousness contributes to the control of purposive activity. This question is distinct from the question of whether some process

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

makes representations conscious, and is therefore a fair demand to make of a vehicle theory.

What can a vehicle theory of consciousness contribute to understanding the function of consciousness? Like some other vehicle hypotheses, O&O's hypothesis suggests that conscious contents are also special in terms of their role in the brain's information processes. For example, MacKay (1990) argued that conscious contents should be identified with long-lasting activation of node structures in his network theory (arguably a vehicle hypothesis), and that the processing function of such long-lasting activation is to allow the strengthening of new links. In MacKay's theory, this idea is said to account for the apparent role of consciousness in representing novelty and acquiring new knowledge. In the target article, stability of activation patterns – and thus phenomenal experience – allows communication among modules. And because phenomenal experience is multimodal, its experienced unity must somehow lie in this communication. The processing function of consciousness is therefore involved in bringing "connectedness and coherence" to the various information-processing activities of the brain.

A critical aspect of O&O's contribution is thus their argument (sect. 5.3) that "subject unity" – what I would describe as the informational specification of self – is compatible with their distributed, modular view of "consciousness making." Their account rests on the temporal and (represented) spatial coherence of distributed brain activity that generates separate phenomenal contents. In their brief development of this notion, it seems that the "phenomenal" status of contents is logically (and perhaps causally) prior to their status as generators of a phenomenal subject. I would argue, however, that these features of experience – the phenomenal status of experience and the generation of an experiencing agent – are of equal and reciprocal theoretical status (Carlson 1997). This view seems compatible with the hypotheses advanced in the target article, and might be viewed as complementary to them.

I see the vehicle theory developed in the target article as an advance in thinking about consciousness and its information-processing support. In particular, I think, developing the kind of theory of conscious agency sketched in the previous paragraphs does depend on an account of informational support that includes a qualitative distinction between the information processing that constitutes conscious contents and that which provides only dispositional representation. Linking degrees of abstractness in phenomenal experience to the hierarchical organization of networks in the brain also seems promising, but it seems that this account depends on the possibility of interaction among levels without phenomenal experience and thus without stable activation patterns, unless all possible levels of abstraction are somehow thought to be simultaneously present in phenomenal experience. Although there may be technical problems with the specific account offered in the target article, I know of no other account that provides this degree of detail in linking properties of informational vehicles to those of phenomenal experience. Such an account will surely be part of a completed theory of consciousness.

Does explicitness help?

Jennifer Church

Department of Philosophy, Vassar College, Poughkeepsie, NY 12604.
church@vassar.edu

Abstract: The notion of an explicit representation plays a crucial role in O'Brien & Opie's arguments. Clarifying what explicit representation involves proves difficult, however, as various explications of this key notion fail to make sense of the overall argument. In particular, neither the notion of encoding in discrete objects nor the notion of active versus potentially active representation seems to help in specifying what is distinctive of conscious representation.

The notion of an explicit representation plays a crucial role in O'Brien & Opie's (O&O's) arguments. It is first introduced as a way to save vehicle theories of consciousness from an obvious sort of objection. Vehicle theories of *x* maintain that it is the stuff that *x* is made of and not what *x* does that makes *x* an *x;* it is the medium and not the message that counts. So, with respect to consciousness, a vehicle theory will hold that it is the material composition (i.e., the neural stuff) of a mental state rather than its relations to other states (i.e., its causal role) that makes it conscious. A vehicle theory may be tempting because of the perceived failure of relational ("process") theories, because it denies consciousness to computers, because it promises more complete reductions, or for any number of other reasons. But any vehicle theory must address the fact ("orthodoxy in cognitive science") that although all of our mental states are made up of brain stuff, not all of our mental states are conscious. One could try to distinguish between different sorts of brain stuff, one being a vehicle for consciousness and the other not, but this is not very promising, given that our brains are composed entirely of neurons and neurons are all pretty much alike.

O&O offer a different solution – one which depends on the notion of explicit representation. Initially, following Dennett (1982), they define explicit representation as representation whereby "each distinct item of data is encoded by a physically discrete object" – in contrast to "information that is stored in a dispositional fashion, or embodied in a device's primitive computational operations" (sect. 1, para. 6). I am not at all sure what counts as a physically discrete *object* in this context – Must an explicit representation be encoded in neurons that encode only that information? Must the neurons be adjacent to one another? Does a unique pattern of firing among neurons count as a physically discrete object? I do not see why information that is "stored in a dispositional fashion" cannot also be encoded by a physically discrete object. Unconscious states, after all, are also encoded (originally or through learning) in particular states of particular neurons. The theory eventually recommended by O&O equates consciousness with "a stable pattern of activation" "somewhere in the brain," where the emphasis seems to be on a stable pattern of *activity* rather than a discrete object (though perhaps the very stability of the activity constitutes it as a kind of object?); but here again I do not see why unconscious states could not be explicit in this sense (after all, they have no causal efficacy whatsoever unless they are in some way active, and O&O themselves say that implicit representations are causally efficacious in their view). So, in the end, I don't see how either explication of explicit representation – in terms of discrete objects or in terms of stable activity – can solve the stated problem with vehicle theories of consciousness.

On the assumption that a vehicle theory should identify consciousness with explicit representation, O&O seek to remove what they view as an obstacle to such a theory – so-called "dissociation" studies (of dichotic listening, blindsight, etc.) that seem to indicate the possibility of explicit representation without consciousness. By reinterpreting these studies, they hope to make room for vehicle theories of consciousness. But it is hard to see why the possibility of dichotic listening or blindsight, in any interpretation, threatens the identification of consciousness with explicit representation. If explicit representation means encoding by discrete physical objects, is there anything about blindsight that suggests that the unconscious representations are not so encoded? Or if explicit representation turns on the stable firing of neurons versus background adjustments in neural propensities ("connection weights"), is there anything about dichotic listening that suggests that the registration of sound must involve stable firing patterns rather than background propensities? As far as I can see, all that is at issue in the dissociation studies is the possibility of receiving and responding to auditory and visual information in the absence of (phenomenal) consciousness, and such causal connections, by hypothesis, are irrelevant to a vehicle theory.

When O&O introduce the contrast between classical and con-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

nectionist models of the mind, the confusion deepens. They maintain that vehicle theories equating consciousness with explicit representation (the only kind of vehicle theory they consider defensible) are not available to classical as opposed to connectionist models of the mind. This is because both conscious and unconscious representations are said to be explicitly represented in the classical model. The information that a cat is on the mat, for example, is encoded in a discrete physical object (a neurally realized symbol) according to the classical model, whether or not it is conscious (sect. 3.1, para. 8). O&O maintain that a connectionist model allows much more information to be encoded nonexplicitly, in the form of structural features of neural networks – for example, connection weights that affect the active firing patterns of a neural network, making the equation of conscious representation with explicit representation much more palatable. But, again, if explicit representation means encoding in discrete objects, I fail to see why certain firing patterns as opposed to certain structural features of neurons that contribute to those patterns count as discrete objects; and if explicit representation means actual versus potential neural activity, I fail to see why the structurally encoded information is not also explicit whenever the relevant neurons are involved.

As is well known, connectionist models of representation may be viewed as either compatible or incompatible with classical computational theories, depending on whether the relevant neural activation networks are thought of as realizations of particular symbols ("cat," "is on top of," "mat"), which are then combined in accordance with various grammatical rules ("the cat is on top of the mat" and "the mat is on top of the cat" are both allowable, but "is on top of the cat the mat" is not) or are thought of as modes of representation that bypass the need for separate representations for separate concepts ("the cat is on top of the mat" represented, as a whole, by one neural network, and "the mat is on top of the cat" represented by quite another).

In the former case, where connectionism is compatible with classicism, neural networks might be thought of as the vehicles of representation, still leaving open the question of whether it is the vehicle or the processes surrounding the vehicle that make it the representation it is. (Does a particular neural network represent a cat in virtue of its causal connections to cats and/or to other neural networks, or does it represent a cat in virtue of its intrinsic features?) Likewise, the stable firing pattern of a neural network might be the vehicle for consciousness, still leaving open the question of whether it is the vehicle or the processes surrounding the vehicle that make it conscious. O'Brien & Opie clearly favor a vehicle theory, but I wonder if it is a representation's stability rather than its explicitness (however this is understood) that really matters in the end and whether such stability is important precisely because stable events are needed for stable causal roles. If so, their understanding of the neurological underpinnings of consciousness may be more compatible with classical theories than they suppose.

What, exactly, is explicitness?

Hugh Clapin

School of Philosophy, The University of Sydney, Sydney, N.S.W., 2006 Australia. **hugh.clapin@philosophy.usyd.edu.au**
www.arts.su.edu.au/Arts/departs/philos/clapin/index.html

Abstract: O'Brien & Opie's theory of consciousness relies heavily on a distinction between explicit activation vectors and inexplicit weight vectors. But determining which representations are explicit vehicles requires appeal to process, and so their vehicle theory is in fact a process theory.

According to O'Brien & Opie (O&O), the difference between explicit (activation vector) and inexplicit (weight vector) representation is just right to explain the difference between conscious and unconscious thought. But the explicit/inexplicit distinction relied on so heavily by O&O has been found sorely wanting in recent careful analyses (Clark 1993; Kirsh 1990), and the only plausible way to draw the distinction to make activation vectors explicit while weight vectors are inexplicit turns O&O's into a process theory.

Contrary to O&O's implication, Dennett's (1982) definition of explicitness does not require that explicit representations be "each possessed of a single semantic value." Cummins (1986) does not make this assumption, nor does Pylyshyn (1984). Certainly the paradigm example of explicit representation is written language; however, Dennett was careful to include a large range of representations as "explicit": "They need not be linear, sequential, sentence-like systems, but might, for instance, be 'map-reading systems' or 'diagram interpreters'" (1982, p. 216).

Thus the notion of "explicitness" relied on by O&O is much stronger than that used in the literature they cite. The more recent literature on this topic (Clark 1993; Kirsh 1990) argues convincingly that there is no clear, absolute distinction between the explicit and nonexplicit representation of information. O&O's notion of explicitness is idiosyncratic and requires significant defence, which they do not provide.

O&O also assume that there is an important distinction between weight and activation vectors, one that grounds the distinction between conscious and unconscious representations. But Dennett's definition of "explicit" allows for information represented in the weight vectors of a connectionist network to be represented explicitly. We can point to a "physically structured object" (the network's interconnections and weights) in the "functionally relevant place" (between the input and output nodes) which holds the information. This object is one of many possible instances of a connected network with a set number of input, output, and hidden nodes and so can be thought of as a "tokening" of a "member" of a general "system" of representation (possible weight vectors), which has a "semantics." The information in the connectivity of the network is used or "read" when there is activity across the input nodes.

Superpositional representation may be explicit. A CD-Rom or vinyl recording of an orchestra represents the sound made by the instruments in the orchestra superpositionally and explicitly. If the contents are the notes played by each instrument, the representation of any given note, chord, or passage equally represents the sound made by each instrument and each instrument's sound is represented by all parts of that representation. If you accept that a CD-Rom represents music explicitly, then you should accept that superpositional weight vectors represent explicitly too.

A related point to notice is that activation vectors may represent superpositionally. It is wrong to claim that "no activation pattern ever represents more than one distinct content" (sect. 4.1, para. 7). Consider a three-layer network whose input units are divided into two subsets which allow two inputs to be presented simultaneously, one input to each subset of input units. In such a network the activation vector corresponding to the hidden units would superpositionally represent the two inputs. Note also that in his careful analysis of superpositional representation, Van Gelder (1991, p. 43) explicitly notes that activation vectors can represent superpositionally. [See also Van Gelder: "The Dynamical Hypothesis in Cognitive Science" *BBS* 21(5) 1998.] So superpositionality does not ground the difference between weight and activation vectors.

Nonetheless, there is *some* difference between activation and weight vectors, and O&O could change their story and claim that whatever that difference is, it underpins the difference between conscious and unconscious knowledge.

What is the key difference? The most plausible candidate is multiple useability. The information represented by activation vectors seems to have a greater range of possible uses in a complex system than does that in the weight vectors – the latter information is not very portable and is not easily available in other parts of the system. This gets at what is plausible in the authors' grounding of consciousness in explicitness: an acceptable theory

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

of consciousness requires that the information we are conscious of be available for many different uses.

Now Clark (1993) argues that knowledge is represented explicitly in proportion to the ease with which it is deployable, and the greater the number of different uses to which it can be put. But it follows from this view that in a possible system which reads and makes use of weight vector information more easily than it does activation vector information, the weight vectors are more naturally thought of as "explicit," and thus in O&O's account, conscious.

Determining whether certain information is represented explicitly or inexplicitly depends on the availability of that information to various processes in the system. If conscious states are explicit states, then the theory of consciousness inherits this dependence.

The bottom line, then, is that the authors' is a process theory of consciousness: a state's explicitness (and thus whether or not it is conscious) depends on the processes it can undergo and its role in the information economy of the system it is a part of.

ACKNOWLEDGMENT

Thanks to John Sutton, Gerard O'Brien, and the Macquarie University Philosophy of Psychology group for valuable discussion.

Stability and explicitness: In defense of implicit representation

Axel Cleeremans[a] and Luis Jiménez[b]

[a]*Cognitive Science Research Unit, Université Libre de Bruxelles CP 122, 1050 Brussels, Belgium;* [b]*Department of Psychology, Universidad de Santiago, 15706 Santiago, Spain.* **axcleer@ulb.ac.be srsc.ulb.ac.be/axc www/axc.html jimenez@usc.es**

Abstract: Stability of activation, while it may be necessary for information to become available to consciousness, is not sufficient to produce phenomenal experience. We suggest that consciousness involves access to information and that access makes information symbolic. From this perspective, implicit representations exist, and are best thought of as subsymbolic. Crucially, such representations can be causally efficacious in the absence of consciousness.

While the hypothesis that information can be causally efficacious (i.e., influence behavior) yet not be available to consciousness is central in most theories of cognition, this assumption has seldom been questioned as directly as O'Brien & Opie (O&O) do in their target article. We must point out right at the outset that it seems so intuitively obvious to us that we can do more than we seem to be aware of that the current widespread questioning about the role of unconscious cognition in general (i.e., Shanks & St. John 1994) is rather puzzling from our perspective. The basic issue is simple: Can knowledge be "in the system" and influence behavior without being available to consciousness? Even though providing an empirical answer to this question has proven far more difficult than previously thought, the theoretical answer has typically been to deny the problem altogether. This is simply because, in the classical framework, representations are passive and only become causally active when they are accessed (interpreted, manipulated, etc.).

From this perspective, therefore, the only possibilities to allow for unconscious cognition consist in (1) assuming that all the relevant knowledge is permanently embedded in the functional architecture, or (2) assuming the existence of a powerful unconscious system that is basically the same as the conscious one, only minus consciousness (see Cleeremans, 1997, for further analysis). O&O successfully defend the claim that classical systems are therefore simply inadequate to conceptualize the implicit/explicit distinction and offer the connectionist framework as an alternative – a perspective that we very much agree with, having defended it elsewhere (see Cleeremans 1997; Cleeremans & Jiménez, submitted). However, O&O then strangely end up claiming, based on superficial review of the relevant literature, (1) that most existing empirical evidence supporting implicit cognition is flawed and (2) that we are phenomenologically aware of any stable activation pattern in our nervous system. In so doing, O&O paint a picture of cognition that again seems to rule out unconscious representation altogether (see also Perruchet & Vinter 1997).

We strongly disagree with these conclusions, while simultaneously espousing the connectionist approach as the framework of choice to understand implicit cognition. The main issue is that by relying so much on stable representations as a vehicle for conscious awareness, O&O's theory ultimately runs into deep conceptual problems. First, the assumption that stability generates phenomenal experience is wholly unsupported: not only do stable patterns exist in our nervous systems that we seem to be incapable of ever becoming directly aware of (e.g., stable patterns of activation over the light receptors of our retinas), but we also fail to be convinced by the argument that stable patterns in artificial systems such as connectionist networks do not generate phenomenal experience because they are mere simulations, not the real thing. In answer to the first point, O&O suggest that plenty of phenomenology falls outside the beam of attention, but this argument only substitutes one problem for another: the "vehicle" theory of consciousness thus appears to be viable only at the cost of requiring a "process" theory of attention.

The second point is likewise problematic in that O&O borrow Searle's (1992) mysterious "causal powers" of real biological systems to defend the claim that artificial stability, in contrast to biological stability, is incapable of producing phenomenal experience. In the absence of supporting arguments, however, this claim is merely a matter of belief. Note that the alternative perspective is no more satisfactory: if indeed one assumes that any stable pattern of activity in an artificial network is sufficient to generate phenomenal experience, the inevitable consequence is that one is then forced to accept panpsychism – a radical step that only few are willing to make (e.g., Chalmers 1996).

In short, stability does not appear to be sufficient to support phenomenal experience. Is it necessary? Let us start by asking how stability is involved in information processing. A basic principle of the connectionist approach (McClelland 1979), namely, cascaded and graded processing, is that a given module can start processing before its inputs have reached stability. In other words, unstable patterns of activation can be causally efficacious, as nicely illustrated by Mathis and Mozer (1995; 1997) or by Becker et al. (1997) through the formalism of attractor networks. Such patterns are no less representational than stable ones: the entire activation space at each layer of a connectionist network is thus both representational and causally efficacious. In such models, stable representations enjoy an enhanced status, and may thus provide the grounds for availability to consciousness, but by the same token, they also are merely specific points in an otherwise similarly causal and representational space. It is thus surprising to see O&O write that "prior to stabilization . . . there is no determinate pattern of activation, and hence no single, physically structured object that can receive a fixed interpretation." Such sentences appear only to restate the assumption, and leave completely unspecified how static a pattern of activation should be in order to constitute a "physically structured object" or worse, who produce their "fixed interpretations." Ultimately, by assuming that *all* patterns of activation are explicit and conscious, O&O end up adopting a position similar in some respects to the classical framework they otherwise reject, and as other commentators point out (Vinter & Perruchet, this issue), this perspective results in deep problems with learning.

O&O's use of the language of process theories of consciousness in this context is also a hint that what makes a representation explicit is not mere stability, and in fact, O&O's interpretation of the term "explicit" varies throughout the paper, by being sometimes taken to be equivalent with causal efficaciousness and consciousness, sometimes with "externally identifiable representation," and sometimes with "interpreted representation" (as opposed to first-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

order patterns). In this respect, Dennet's definition of explicitness as quoted by O&O appears to us to be the most convincing: information is explicitly represented whenever it involves a pattern that is being interpreted by the system as a representation – a symbol. Crucially, this definition does not rule out causally efficacious sub-symbolic *representations,* which is exactly what we believe to be necessary to understand implicit cognition.

The stability criterion may therefore be a necessary condition to support consciousness, but it does not appear to be sufficient to support metaknowledge, that is, to support a form of consciousness that involves access to knowledge *as* knowledge. This form of consciousness, however, is crucial for abstract thought and explicit learning, and is probably what specifically characterizes human cognition (Clark & Karmiloff-Smith 1993). Hence in our view patterns of activation in connectionist networks are continuously causally efficacious, whether stable or not, and do not in and of themselves generate phenomenal experience. Rather, they are potentially available to consciousness depending on other factors such as stability, strength, global coherence, access by some other structure, or their compositional and systematic character. Such patterns are best characterized as subsymbolic. The genuinely hard problem is then to determine how such patterns can become symbolic, explicit, and conscious, that is, how they can be taken by the cognitive system *as representations.*

ACKNOWLEDGMENT

Axel Cleeremans is a Research Associate with the National Fund for Scientific Research (Belgium).

Trains, planes, and brains: Attention and consciousness

Max Coltheart

Department of Psychology, Macquarie University, Sydney NSW 2109, Australia. **bhs@mq.edu.au**

Abstract: O'Brien & Opie believe that some mental representations are evoked by stimuli to which a person is attending, and other mental representations are evoked by stimuli to which attention was not paid. I argue that this is the classical view of consciousness; yet this is the view which they wish to challenge.

Suppose you are a trainspotter, squinting through your binoculars across the Thames at a train, and you have just identified it definitely as a Class 156 Sprinter and not a Class 143 Pacer. So, O'Brien & Opie (O&O) would say, a neural net has achieved a stable state (the 156 Sprinter State) and avoided another (the 143 Pacer state).

Suddenly you notice through your binoculars a small black spot in the sky just above the train, rapidly looming larger; soon, from its exotic appearance and the surprising lack of noise, you identify it as a Grumman B-2 Spirit (you're a planespotter as well) – which means that another neural-net stable state has been achieved. Now, what happened to the Sprinter stable state? Is it still present? Can two different stable states coexist?

O&O answer in the affirmative: "Connectionism . . . treats the mind as a large collection of interconnected, specialized PDP networks, each with its own connectivity structure and potential patterns of activity. . . In other words, according to connectionism, from moment to moment the brain *simultaneously* [my emphasis] realizes a large number of explicit representations" (sect. 5.2, para. 1). Each of these simultaneous stable activation patterns is identical to a phenomenal experience.

With respect to the example with which I began, not only are there two distinct stable states representing the train and the plane, but also, since O&O's view is that the presence of a stable state evoked by an object is identical to being conscious of the object, the person in my example is simultaneously conscious of the train and the plane, according to O&O. That does not accord with my experience, which tells me that I am only ever conscious of one thing at a time.

The authors attempt to address this objection in their footnote 26: adapted to my example, their account of this situation is that the observer is equally conscious of the train and the plane, the difference being that he is attending to the plane but not attending to (though still conscious of) the train. In that case, of course, there is an urgent need for an account of the distinction between attention and consciousness in their connectionist framework. If two stable states exist, one evoked by the train and one by the plane, what is it about these states that distinguishes the one that is enjoying the observer's attention from the one which is not?

O&O attempt the following solution:

> 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 . . . to more intense processing. Such additional processing would engage 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. (Note 26)

This to me seems unsatisfactory for several reasons. First, there is the phenomenological point that the difference between the observer's experience of the train and his experience of the plane is not merely quantitative, but qualitative. Second, whilst O&O agree in Note 2 that "attention is clearly restricted to a single focal object at a time," their connectionist account of attention does not require this; why could not two different stable patterns of activation engage extra neural networks to augment phenomenal experience, which would mean that attention to two different focal objects would be occurring? One answer might be: because this extra system of neural networks needed to bestow attention can only be used in relation to one stable pattern at a time. But now the vehicle theory collapses back into a process theory. O&O define such a process theory as one in which

> our conscious experience is the result of a superordinate computational process or system that privileges certain mental systems over others. . . 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 specific kinds of computational processes. We shall call any theory that adopts this line a *process* theory of consciousness. (sect. 1, para. 7)

On this definition of process theory, their theory of attention is a process theory, not a vehicle theory.

Given what they say about attention, O&O are arguing that there are two types of mental representation: both are by definition conscious (because they argue that all mental representations are conscious), but one type is attended-to and the other is not. In what way is that different from the standard view in cognitive science, according to which there are two types of mental representation, one conscious and one unconscious? O&O's unattended-to mental representations are identical to the standard theory's unconscious mental representations (because on the standard theory the mechanism that converts an unconscious representation to a conscious one is, precisely, attention).

This leads us to the issue of subception – subliminal or unconscious perception. On O&O's view, there can be no mental representation without conscious experience. Hence they are moved to review certain bodies of literature purporting to show that there can be unconscious mental representations – literature on dichotic listening, blindsight, implicit learning, and visual masking – and to conclude that the empirical evidence for dissociation between mental representation and conscious experience is not strong. But for all of the work they review, the classical view of consciousness could just as well describe these results as show that there can be dissociations between attention and mental representation – that a stimulus to which the observer did not attend nevertheless can evoke mental representations. And O&O would not want to dispute this – on the contrary, they affirm it (in their

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Note 26). They should therefore be perfectly happy with the idea that when a person is instructed to attend only to the stimuli in the right ear, stimuli presented to the left ear can nevertheless evoke mental representations. Why, then, do they wish to challenge the conclusions from studies of dichotic listening and selective attention?

In sum, I see no difference between O&O's claim (there are two kinds of mental representations, the attended-to and the not attended-to, and the classical claim (there are two kinds of mental representation, the conscious and the unconscious). What difference does it make whether you say "at that point, he was conscious of the plane but not of the train" or "at that point, he was attending to the plane but not to the train"?

Stability is not intrinsic

D. C. Dennett and C. F. Westbury

Center for Cognitive Studies, Tufts University, Medford, MA 02144.
ddennett@tufts.edu cwestbur@emerald.tufts.edu
www.tufts.edu/as/cogstud/mainpg.htm

Abstract: A pure vehicle theory of the contents of consciousness is not possible. While it is true that hard-wired tacit representations are insufficient as content vehicles, not all tacit representations are hardwired. O'Brien & Opie's definition of stability for patterns of neural activation is not well-motivated and too simplistic. We disagree in particular with the assumption that stability in a network is purely intrinsic to that network. Many complex forms of stability in a network are apparent only when interpreted by something external to that network. The requirement for interpretation introduces a necessary functional element into the theory of the contents of consciousness, suggesting that a pure vehicle theory of those contents will not succeed.

One can be grateful for a theory such as the one offered, without being convinced by it, since O'Brien & Opie (O&O) resolutely explore some tempting but foggy territory. If our verdict about their exploration is negative, at least now we may be able to see clearly for the first time why we were wise to sidestep this option.

O&O's criticisms of the prevailing assumptions about unconscious information processing are timely and important. Although we have some minor quarrels with some of them, we agree that the standard assumption that there is a sharp (or principled) distinction between unconscious and conscious information-processing is misbegotten. They say: "it is not unreasonable to reserve judgment concerning the dissociability of explicit mental representation and phenomenal experience" (sect. 2.4, para. 4). We would put it somewhat more strongly. This oft presupposed dissociability depends on distinguishing between unconscious information processing on the one hand and very brief intervals of conscious-but-soon-forgotten information processing on the other, and this is not supportable. It presupposes what Dennett (1998) has called the myth of double transduction: the idea that unconscious contents in the brain become conscious by being transduced into a privileged neural medium (as most clearly expressed by Mangan 1993a; 1996).

The well-named "classical" approaches to cognitive science (whose name hints that they belong behind glass in a museum somewhere) do indeed propel the theorist headlong towards the myth of double transduction, but it is not clear that a pure vehicle theory can avoid equally ominous impasses in other directions. We see three related problems. The first concerns a missing taxon in O&O's representational taxonomy, the second their definition of stability, and the third the role that stability of component networks might play in a larger meta-network.

Transient tacit representations: As O&O point out, "hardwired" tacit representations can hardly serve the purposes of content vehicles in any theory of the fleeting contents of consciousness. However, they do not consider the question of whether all tacit representations are hardwired. They are not. Dennett's taxonomy of styles of mental representations includes one further taxon which they overlook, transient tacit representations (Dennett, 1982, p. 224, reprinted in Dennett, 1987, pp. 213–25), which are available for a system's use only when that system is in a particular state. These representations are obviously the most important for the purposes of the argument presented. Indeed, the stable connectionist patterns championed by O&O are presumably just such sorts of mental representations – they call them non-explicit. Although O&O claim that the distinction between potentially explicit and tacit lapses "for all practical purposes," they are thinking only of hardwired, nontransient tacit representations. With transient tacit representations, that distinction is not simply of practical insignificance, but theoretically unmotivated.

The definition of stability: the idea that it is the most influential transient representations in cortical networks that earn the status of consciousness is fine. However, we do not see that O&O have succeeded in defining stability or its influence on the larger cortical network in such a way that one can assess their claim that "only stable patterns of activation are capable of encoding information in an explicit fashion in PDP systems" (sect. 5.1, para. 4); hence we also cannot assess their claim that it is all and only these stable patterns that are vehicles of conscious content.

One problem is simply that it is arbitrary and simplistic to declare that a network is stable if its constituent neurons are firing simultaneously and at a constant rate. Such a simple definition of stability ignores the fact that stability can manifest itself in a network in many more complex ways. Since a network can cycle through time, it can have a (possibly very complex) temporal stability that is impossible to discern spatially because it has no simple spatial representation at shorter time scales than the time it takes to cycle. Such complex stability can be discerned by an entity (including another network) which samples it at the right location and frequency. This idea of complex forms of stability was suggested by Hebb (1949) when he first described his Hebbian cell assembly, which is precisely the mechanism being described in this paper as the holder of phenomenal experience.

A further complication is added if we grant that the sampling system might have the ability to quantize states in the sampled system – that is, to pull information to its nearest category, as a basin of attraction in a complex system equates a wide number of states by the fact that they all lead to the same attractor. It is easy to imagine a network sampling a number of arbitrary points from another network and finding them stable because of its (the sampler's) characteristics, even though there is nothing in the sampled state that shows the stability. Stability is as much a function of the sampler as of the sampled. In a complex system, states that are not empirically identical can be functionally identical. We doubt that defining stability as simultaneous, constant firing will suffice to explain the behavior of myriads of interacting networks in the brain, and we are baffled by the suggestion that stability of the requisite sort is not to be found in serial simulations of connectionist networks – as if the stability of a virtual machine were any less powerful a feature than the stability of an actual machine.

The role of stability: finally, O&O's claim that it is a virtue of their vehicle theory that it makes phenomenal experience an "intrinsic, physical intranetwork property of the brain's neural networks" (sect. 5.1, para. 10) is, we think, confused. If the "intrinsic" property of stability is also an "intranetwork property," then presumably it is the role of the component networks in modulating the larger activities of the entire cortical metanetwork that mark them for the role of phenomenal experience, not their "intrinsic" stability. If it turned out, for instance, that there was a subclass of stable patterns in the networks that did not play any discernible role in guiding or informing potential behavior, would their stability alone guarantee their status as part of phenomenal experience? Why?

Dennett (1991) stressed the importance of this when he proposed what he called "the Hard Question" (p. 255): "and then what happens?" (see also And then What Happens?, Dennett, 1991, pp. 263–75). An instance of the failure to appreciate this

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

point appears in O'Brien & Opie's suggestion that "when phenomenal properties coincide temporally, . . . this is a consequence of the simultaneity of their vehicles" (sect. 5.3, para. 6). The "intrinsic" simultaneity of vehicles could not by itself account for subjective simultaneity. As we have stressed above, what matters is not actual ("intrinsic") simultaneity, but either the (correct or mistaken) detection of simultaneity by the larger system of which these vehicles are a part, or else the failure of the larger system to generate any complaint about their nonsimultaneity. If such functional effects are as vital as we suggest, a pure vehicle theory of consciousness cannot succeed.

Consciousness, connectionism, and intentionality

Donelson E. Dulany

Department of Psychology, University of Illinois, Champaign, IL 61820.
ddulany@s.psych.uiuc.edu

Abstract: Connectionism can provide useful theories in which consciousness is the exclusive vehicle of explicit representation. The theories may not, however, handle some phenomena adequately: sense of agency, modes and contents of awareness, propositional and deliberative thought, metacognitive awareness and consciousness of self. They should, however, be useful in describing automatic, activational relations among nonpropositional conscious contents.

Connectionism has become a powerful intellectual force, and consciousness is the most significant intellectual challenge facing a number of disciplines. Whatever the limitations of a connectionist approach to consciousness, O'Brien & Opie (O&O) can be congratulated for bringing forward a provocative and useful thesis for examination.

Most significant, I believe, is their case that connectionist metatheory provides for consciousness being an *exclusive* vehicle of explicit representation. The view is sure to be challenged by those with a deep commitment to a cognitive unconscious that symbolically represents and can act as an independent system. That assumption has been fundamental to standard architectures, from Bower (1975) to Baars (1996). It is also embodied in computational views (e.g., Jackendoff 1987) holding that consciousness is only a sometime, nonobligatory emergent of fully formed cognitions. O&O are right to describe serious challenges to that view, but no single paper can undo decades of conceptual confusion and methodological bias in the "dissociation" literatures. Suffice it to say now that if claims for the power of a cognitive unconscious were correct, the experimental effects would be too strong and replicable for these literatures even to be controversial. No one can claim that.

If we challenge the dissociation thesis, a major value of connectionist metatheory lies in consigning the nonconscious to mental operations and the potentially explicit; the explicit representations and their forming operations become inseparable within the network. This is inconsistent with symbols in an unconscious being essentially like symbols in consciousness, just "stored" but separately doing what symbols do while waiting to be "retrieved."

But does connectionism provide the computational resources needed for a broad enough vehicle theory of consciousness? And does O&O's conception of "explicit representation" capture what we need to recognize in phenomenal experience?

We can all consciously symbolize something past in remembrance, something present in perception, or some possible future in an intention or expectation – or even something entirely unreal in imagery. If a connectionist theory of phenomenal experience is to be successful, it must handle those phenomenal realities. Indeed, O&O wisely challenge a "classical" paradigm rather than a "symbolic" paradigm, and they essentially argue that connectionist networks relax into symbolic (explicit) representations. Nevertheless, drawing upon Dennett (1982) for a conception of "explicit representation" as a "physically discrete object" does more to satisfy physicalist commitments than to distinguish what is to be explained from a stapler or a BMW. How symbols "explicitly represent" is functionally specifiable in theories, and a mental event is a symbol if it does what theory says symbols do. They evoke other symbols. They can be used in propositional beliefs, including the special one that "this stands for that." And they are contents of mental activity bounded by sensory and motor transducers that permit interaction with the world in ways that warrant those special beliefs.

Once we move beyond conscious experiences that are simpler and automatically evoked – that are nonpropositional and nondeliberative – the connectionist metatheory of consciousness leaves rather vague promissory notes – in this paper, the appeal to a "multitude" of networks (sect. 5.2). These promises are difficult to fulfill gracefully, however, with computational principles that are exclusively associative – as we learned from an earlier "connectionism," S-R theory (Dulany 1968; Thorndike 1949).

1. I see nothing in one or many networks that captures the sense of *agency,* of possession, with which we hold our conscious modes and contents – a property clinicians see diminished in neurosis and sometimes absent in psychosis.

2. Neither do I see principles that would distinguish and combine the outputs of networks so that *contents* of experience would be carried by their appropriate *modes* – of perception, belief, intention, fear, or hope.

3. Even if elements of a proposition should be represented by hierarchical modules, as in McClelland (1995) and Hummel and Holyoak (1997), convergence of their output activations misses strength of belief, the vehicular mode of the proposition. Associative activation from "John," "loves," and "Mary" does not capture the degree of belief that "John loves Mary." Believed predication is not free association.

4. We can see why once we recognize that deliberative operations go beyond associative activation/inhibition. Some mental episodes – call them "evocative" – do constitute a simple perception, or feeling, or sense of one thing activating another. These are nonpropositional contents carried by nonpropositional modes of awareness. But in deliberative episodes, inferences or decisions operate on propositional contents and modes to yield others. We can, for example, believe or disbelieve with certainty on a single piece of convincing propositional evidence (Carlson & Dulany 1988). Deliberative warrant is not association strength.

5. The limitations of associative principles continue to be apparent when the authors venture explanations of metacognitive awareness and a coherent sense of self (sect. 5.3). Meta-awareness derives not only from memory of conscious states but also from deliberative inference as to the forms of mental episodes. And although O&O appeal to "narratives" for a continuing sense of self, the narrative would require a sequence of propositions and a personal sense of agency for the protagonist.

6. Furthermore, between the boundaries of transduction are mental episodes that start as well as end with conscious states – for example, the conscious perception that evokes a remembered past or an imagined future. Conscious states as causal remain outside O&O's approach unless conscious states can be identified with input units as well, perhaps fed from output of other networks. As only the stable states of relaxed networks, consciousness still rides the back of the bus.

It is in taking on more difficult challenges to the connectionist metatheory that the limitations of connectionist modeling have been most clearly revealed (e.g., Green 1998; Massaro 1988; McCloskey 1991). Furthermore, although O&O wisely disavow a claim to solving the "hard problem" of consciousness (sect. 5.4), this undercuts a strong interpretation of their central thesis that "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" (sect. 5).

What this leaves is the promise of connectionist metatheory and

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

modeling for the automatic activation of nonpropositional conscious contents in simpler, evocative mental episodes: transduction of literal awareness (of form, color, pitch, etc.), activation of identity awareness by literal awareness, and associative learning and remembering of novel associations among nonpropositional conscious contents. But a broader program would also include analyses of agency, modes, propositional contents, deliberative episodes, and metacognitive awareness. Together, these constitute an analysis of the *intentionality* of consciousness – one in which I have also proposed that consciousness is the sole carrier of symbolic-explicit representations, with the nonconscious consigned to mental operations and memories that are only dispositional (Dulany 1991; 1997). For this more limited role, O&O's approach would still be a significant contribution.

A note on imaginability arguments: Building a bridge to the hard solution

Ralph Ellis

Philosophy Department, Clark Atlanta University, Atlanta, GA 30314.
rellis@cau.edu

Abstract: According to "imaginability arguments," given any explanation of the physiological correlates of consciousness, it remains imaginable that all elements of that explanation could occur *without* consciousness, which thus remains unexplained. The O'Brien & Opie connectionist approach effectively shows that *perspicuous* explanations *can* bridge this explanatory gap, but bringing in other issues – for example, involving biology and emotion – would facilitate going much further in this direction. A major problem is the ambiguity of the term "representation." Bridging the gap requires perspicuously explaining not just how we form "representations" in the sense of outputs isomorphic to what is represented, but also what makes representations conscious; I sketch briefly what this would entail.

If we can imagine X without Y, does this mean that an explanation of X cannot adequately explain Y? According to O'Brien & Opie (O&O), the real issue is whether explaining X provides some *perspicuous facilitation of understanding* for Y; for example, someone ignorant of chemistry can imagine H_2O without water, but H_2O is not therefore nonidentical with water. Explanations of H_2O are satisfying explanations of water, leaving no explanatory gap, because we understand *how* H_2O yields the familiar properties of water. Whether we can imagine X without Y often depends on our knowledge about X and Y. If we really knew much about Clark Kent, we could not imagine Kent not being Superman, because we would know of evidence that he *is* Superman; this evidence would facilitate understanding *how* a being with *all* the properties of Kent *must* also on other occasions exhibit properties of Superman.

O&O develop just such a perspicuous account of the brain-consciousness relation. Their account is more perspicuous than most because their connectionism demonstrates concomitant variations in brain processes isomorphic with variations in the various kinds of conscious experience, so that we can understand just how the differences *between* sense modalities such as color and taste are reflected in different patterns of connectionist weights, while simultaneously leaving just the right kind of theoretical space for the variations we find *within* each sensory modality.

It is quite possible, however, that this connectionist system only explains how a system of "representations" could be carved or written *in* the brain (or in its functions), without addressing why the written information should be any more *conscious* than, say, a word written across my forehead, or a holographic image floating in the vicinity of my head. The target article leaves connectionism vulnerable in principle to the same problem as traditional computational theories: even a model that simulates the *processing of information* accomplished by conscious beings may not incorporate the feature of consciousness, because the model has "represented" the information in a non-conscious *way* – for example, by means of patterns of electrical circuits yielding outputs, which we know can be accomplished *without* consciousness. A connectionist system yields outputs more similar to conscious and human brains' outputs, but that still does not show that the model incorporates the elements of consciousness, because the information is still being "represented" in a way that can be accomplished without consciousness, just as can the "representation" of information in classical computational models; the connectionist system simply "represents" the information (in this nonconscious sense) *more accurately* (where "accuracy" means degree of similarity to the brain's style of processing). A similar problem occurs with Pribram's holographic theory: even if the brain *processes information* holistically, yielding outputs resembling holograms, to explain how a hologram appears in my head no more explains consciousness than explaining how a conventional photograph inserted into my head would somehow become a *conscious* image just because it is in my head rather than a photo album.

Perspicuity of explanations is the crucial requirement for bridging the explanatory gap; the imaginability of X without Y depends on the degree of knowledge/ignorance of X and Y; but what is needed is a perspicuous explanation that does not equivocate "representation." We need a physical explanation that allows us to understand how the system would give rise to a *phenomenal* sensing of various kinds of imagery, feelings, concepts, and so on, not just "representations" qua information *output* isomorphic to outputs of conscious brains. What else is needed?

Full occipital processing of visual signals does *not* yield perceptual *consciousness* unless frontal and limbic brain areas are also involved – that is, just the brain areas associated with the emotional motivation to look for organismically relevant stimuli (Aurell 1989; Ellis 1995). The *emotional motivations* of biological organisms (on which O&O are completely silent) are necessary aspects of the "felt" dimension of phenomenal experience. For example, subcortical emotional areas of the brain are necessary for perceptual consciousness; the motivational selection of perceptual data for processing, through corticothalamic loops involving limbic and frontal areas such as the anterior cingulate (Posner & Rothbart 1992), is a necessary aspect of conscious attention. Motivatedly looking for a given organismically relevant category of input (involving frontal and limbic areas) gives rise to mental images, with or without incoming afferent signals; this dimension of representation is the one that determines whether the representation is a conscious one. The biological/motivational dimension is needed for a truly perspicuous explanation of why some instances of "representation" are *conscious* while other "representations," even in the brain (e.g., in blindsight) are not. This motivational/biological dimension in no way contradicts the connectionist account, but would add a crucial ingredient without which we still do not have an explanation of conscious representation, but only of "representation" *isomorphic* to our conscious experience.

Network stability and consciousness?

Daniel Gilman

Laboratory of Neuropsychology, National Institute of Mental Health, Bethesda, MD 20892. **dan@ln.nih.gov**

Abstract: A connectionist vehicle theory of consciousness needs to disambiguate its criteria for identifying the relevant vehicles. Moreover, a vehicle theory may appear entirely arbitrary in sorting between what are typically thought of as conscious and unconscious processes.

O'Brien & Opie (O&O) claim to have a simple empirical hypothesis regarding conscious experience; that is, "phenomenal experience consists of the explicit representation of information in neurally realized parallel distributed processing (PDP) networks." Perhaps the hypothesis is not so simple. Without a noncontrover-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

sial theoretical explication of "phenomenal experience," it may be difficult to establish whether phenomenal experience indeed correlates with such explicate representation. Moreover, the notion of explicit representation itself may be problematic. The explicit/implicit representation distinction in neural processing is notoriously difficult (Crick & Koch 1995). O&O stipulate an identity between explicit representation and stable activation patterns in neural networks. But if their account of network stability is unclear, empirical investigation of the psychological properties of stable networks becomes a difficult proposition. I will first consider the clarity of their stability criterion and then turn to the question of the physiological and phenomenal plausibility of a candidate sense of stability.

It appears that O&O have in mind at least two distinct senses of "stability," and that each is problematic.

Sense 1: Neural plasticity: a stable activation pattern is the product of a "settled" network, as evidenced by characteristic, or consistent, output, for a given input.

Sense 2: A stable network is one in which the constituent neurons are firing simultaneously at a constant rate.

Neither sense is especially clear without a relevant time scale and this question is not addressed, except with the contention that many stable states may occur per second.

I begin with Sense 2. Part of the general discussion of connectionism certainly implies that spike count simpliciter is the relevant variable, but again we need a relevant time scale to make sense of the contention. Perhaps much of the information in a spike train is carried by the total spike count, defined over some interval. But not all, apparently, and there are important questions about the information carried in temporal patterns within the spike train (Gawne et al. 1996; Lestienne 1994) and about resolution limits within which timing does not convey information (Heller et al. 1995). These point to significant and unanswered questions about neural coding, questions which may not have a simple answer across all brain regions and types of neurons. Moreover, on this construal of stability, "down" networks, engaged merely in background firing at characteristic background firing rates, look to be reasonably stable and thus engaged in the realization of explicit representations and hence conscious thoughts, though general understanding has them essentially unoccupied. Issues to do with the information bearing significance of how a spike train is played out are by no means trivial. But resolving such issues is, or ought to be, critical to the present discussion.

Sense 1 prompts the obvious question of how settled a network has to be. It is possible, to be strict to a degree that excludes most or even all brain processing. On the other hand, on a liberal construal of stability, the entire brain is stable. This should not be surprising in itself, since a general sense of stability might be taken as equivalent to intelligibility (in this case, across networks) and thus to proper brain function, whatever the brain is doing. Moreover, O&O are clear that they are after a stability/instability distinction, not a slow/transient network distinction, and they point out that many stable states may succeed each other per second. But the idea of maximally settled networks – however transient – suggests we consider very early centers of perceptual processing, reflex networks, and the like to be excellent candidates for consciousness. Fast automatic mechanisms are not typically thought to instantiate conscious states, but they may be excellent exemplars of consistently behaving networks. Crick and Koch (1995; 1998) have rejected the notion that consciousness arises as early as V1 in the ventral processing stream. Their suggestion may be contentious. But are we to take it that consciousness may arise not only in V1 but in LGN? In the retina? Is this "phenomenally plausible," given what we know of the retinotopic – not object-centered – response properties of very early perceptual networks? Likewise for mechanisms of physical and psychological reflex. We are told that "understanding experiences, in particular . . . presumably corresponds to stable patterns of activation in higher-order networks . . . and are thus . . . most subject to voluntary control." (sect. 5.2, last para.) How well does this sit with their paradigm example of an understanding experience, understanding one's own language when one hears it? Is this most subject to voluntary control? Finally, internetwork properties are not held to be important and neither is environmental embeddedness. Is a 28 cell network, stable in vitro, conscious? If so, of what?

ACKNOWLEDGMENT

This comment was written with the support of a National Science Foundation Professional Development Fellowship.

When is information represented explicitly in blindsight and cerebral achromatopsia?

R. W. Kentridge

Psychology Department, University of Durham, Durham DH1 3LE, United Kingdom. **robert.kentridge@dur.ac.uk** **www.dur.ac.uk/~dps1rwk**

Abstract: Discrimination of forms defined solely by color and discrimination of hue are dissociated in cerebral achromatopsia. Both must be based on potentially explicit information derived from differentially color-sensitive photoreceptors, yet only one gives rise to phenomenal experience of color. By analogy, visual information may be used to form explicit representations for action without giving rise to any phenomenal experience other than that of making the action.

I am largely sympathetic to O'Brien & Opie's (O&O's) position that consciousness might be built from many components and that consciousness of each of these components depends on their explicit representation. I do not, however, see that this model is necessarily inconsistent with neuropsychological dissociations between consciousness and performance.

O&O refer to a number of neuropsychological conditions in their article. In particular, they cast doubt on the dissociation between consciousness and performance in blindsight, based on the arguments of Campion et al. (1983) and they refer to the independence of processing modules as revealed by neuropsychological deficits of motion or color perception (Sacks 1985; Zeki 1993; also see Meadows 1974 and Zihl et al. 1983 as primary sources). The criticisms of the phenomenon of blindsight raised by Campion et al. (1983) were well dealt with at the time in commentaries, but since then further evidence has amassed that residual visual function in blindsight cannot be explained by scattered light (King et al. 1996), changes in subjects' decision criteria (Azzopardi & Cowey 1997; Kentridge et al., in press) or islands of spared visual cortex (Kentridge et al. 1997; Stoerig et al. 1998). O&O characterize blindsight as a subcortical phenomenon; they may not be right in this. Anatomically, it is possible for visual information to reach the cortex via projections that bypass striate cortex (see, e.g., Stoerig & Cowey 1995). Evidence from a recent functional magnetic resonance imaging study suggests that although visual stimuli do not elicit any activation in the damaged striate cortex of a blindsight subject, they do produce activation in extrastriate cortical areas (Stoerig et al. 1998).

This need not affect O&O's position if they are right in their suggestion that although blindsight subjects do not have normal visual experience in their blind fields, they do have other types of phenomenal experience associated with visual stimuli.

The extent to which residual visual ability and consciousness are dissociable after lesions to striate cortex is, therefore, still an issue crucial to O&O's position and worthy of further discussion. O&O are right that subjects with residual vision in scotomata caused by striate cortex lesions often report some awareness of events in their blind regions. If, however, subjects are asked to report, on a trial by trial basis, whenever they have any such experience whatsoever, there is a clear dissociation between their residual visual abilities and their awareness (see, e.g., Weiskrantz et al. 1995; Kentridge et al. 1997; Zeki & ffytche 1998). In other words, there are conditions in which these subjects do not report any phenom-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

enal experience associated with a stimulus, yet are still able to make a correct explicit judgment about its nature. Does this inevitably lead us to some form of executive or process theory of consciousness or can a vehicle theory survive?

O&O's model implies that any cortical stimulus representation that is explicit, in that it supports stimulus specific behaviors, must also give rise to consciousness. I have noted that this does not appear to be the case in blindsight. There are many other neuropsychological examples. In cortical color blindness subjects not only deny conscious experience of the hue of stimuli, they are incapable of discriminating hues in forced-choice tasks (Heywood et al. 1998). Nevertheless, when presented with a stimulus consisting of a figure and background that differ from each other only in color (i.e., they are equiluminant), they can effortlessly (and consciously) see the figure (Barbur et al. 1994; Heywood et al. 1994). Color information that is present but presumably not explicitly represented in the subjects' undamaged brain regions could give rise to an explicit and conscious representation of form without giving rise to conscious experience of color.

O&O note that different brain regions subserve different functions (for example, motion perception) and that these functions can be computed on the basis of a variety of different types of stimuli (motion might be perceived on visual or auditory stimuli, for example). In my cortical color blindness example we might suppose that the module capable of extracting form from color differences (or, more specifically, from differences in the activation of cells responsive to specific wavelengths of light at different points in space), luminance differences, texture differences, and so on, is intact, whereas the module in which hue is extracted from ratios of wavelength specific activations at common points in space (a quite different calculation) is damaged. In other words, the second module depends on cells that are selectively activated by light with specific spectral content, while the first may utilize differences in the responses of color sensitive (but not color selective) cells across a color border: these cells signal color variation without coding the nature of the hues which comprise the border. Hue may be potentially explicit in both of these modules, but the structure of the network only makes it explicit in the second.

I hope O&O will have no problem with this distinction between potentially and actually explicit representations in cerebral achromatopsia. Can blindsight be dealt with in a similar manner? The activities of cells driven by pathways originating in the eye depend on the way in which cells in the retina respond to the number and wavelength of the photons which fall upon them. Extracting the surface properties of objects like brightness or texture from this signal requires computation just as computations have to be made to extract hue from a number of different wavelength sensitive activations. Is there any reason to believe that all the modules that process this visual signal give rise to visual experience? The extraction of form from color does not give rise to a color experience. The dorsal stream leaving the primary visual cortex and passing up to the parietal cortex is strongly associated with action based on visual stimuli (Milner & Goodale 1996). Is it unreasonable to suggest that components of the visual signal that are potentially explicit elsewhere are extracted in this stream and give rise to action (and an awareness of action) without giving rise to a visual experience?

The gap into dissolution: The real story

Martin Kurthen

Department of Epileptology, University of Bonn, D-53105 Bonn, Germany.
martin@mailer.meb.uni-bonn.de

Abstract: For a theory of phenomenal consciousness, the real issue is not that between vehicle and process, but between naturalistic and deconstructive theories. Most current naturalistic theories combine a hypothesis about the neural correlate of consciousness with a subsequent naturalistic proposal about how to close the explanatory gap. Deconstructive theories use theses about the neural correlate of consciousness only to motivate and support their claim that the "hard problem" of consciousness is a pseudo-problem which is not to be solved, but rather *dis*solved on nonnaturalistic grounds. O'Brien & Opie present a hypothesis concerning the neural correlate of consciousness, but no genuine strategy to close the explanatory gap. Their theory can, however, contribute to the success of a deconstructive theory of PC.

The persistence of the explanatory gap in a connectionist theory. The core of most naturalistic theory of phenomenal consciousness (PC) is a hypothesis about the correlate. But proponents of naturalism can choose between a number of different stances: they can hold that the correlate will somehow be identified with phenomenal consciousness, that knowledge of the correlate will somehow change our view on or our concept of PC, that owing to our cognitive dispositions we are unable to grasp the way in which phenomenal consciousness is realized as a natural phenomenon, and so on. In their excellent target article on vehicle theories of phenomenal consciousness, O'Brien & Opie (O&O) state that their connectionist vehicle theory of phenomenal experience, which is better categorized as a connectionist theory of the correlate is "no worse off . . . than any other current theory" (sect. 5.4) with respect to the hard problem of why the brain gives rise to phenomenal properties at all. But they hold that the vehicle theory may change our view on the property of having phenomenal consciousness: "If we can find a neural mechanism that mirrors in a systematic fashion the complex structural properties of phenomenal experience, it may eventually be inconceivable that a creature with this mechanism would not be conscious." (sect. 5.4).

This argument is misleading: (1) It is by no means clear that the neural correlate of a phenomenally conscious state should have to be analogous or similar to that state at all. For example, the correlate of a unitary experience might well be a spatially and temporally dispersed and discontinuous pattern of neuronal activity (or, if it could not be, this would have to be demonstrated independent of the mere intuition of analogy or even aesthetic adequacy). (2) Even if such a "neural mechanism" could be determined, the conceivability of the occurrence of that mechanism without an accompanying phenomenal consciousness would not be ruled out. (3) Even if the state of unconsciousness were inconceivable in the face of the mechanism in question, the cooccurrence of the two relata (phenomenal consciousness and the analogically structured mechanism) could neither explain the internal constitution of the relata themselves (which is taken for granted) nor the nature of their relationship, which has to be interpreted as identity, supervenience, or whatever, on the basis of an independent justification (Kurthen 1995). All this holds for connectionist and classicist, vehicle and process theories alike, because it holds for any theory that offers an allegedly plausible account of the neural correlate but not more.

Hence, the dichotomies of connectionism versus classicism and vehicle versus process theories may well be crucial for a theory of the correlate, but they have no direct significance for the hard problem in the theory of phenomenal consciousness. Furthermore, connectionist and/or vehicle theories are not generally superior to classicist and/or process theories with respect to the problem of how to close the explanatory gap, although they may yield a more plausible account of the correlate (which O&O seem to argue for when they find that – not least due to the rise of connectionism – vehicle theories deserve some rehabilitation). A theory of the correlate is one possible starting-point for a theory of phenomenal consciousness, but the closing of the gap is not part of the same theoretical project as the demonstration of the correlate. But although the vehicle theory will not close the gap by means of a better description of the correlate, it may still contribute to a theory of PC by supporting a deconstructive strategy towards PC.

Deconstruction of PC as a fruitful strategy. According to deconstructive theories of PC the hard problem is a pseudoproblem. The deconstructionist holds that the only genuine problem for a

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

scientific theory of phenomenal consciousness is determining a correlate: deconstruction and dissolution will do the rest. The typical naturalist presents a hypothesis concerning the relation between PC and its neural correlate which, if it were true, would enable us to naturalize PC – the identity thesis is a proper example for this. The deconstructionist, on the other hand, argues that we are already wrong when we try to find such a hypothesis; in his view, there *is* nothing to explain (although there may seem to be). To me (Kurthen et al. 1998), the most natural deconstruction of phenomenal consciousness follows from Rorty's (1993) proposal that consciousness is relational or description-dependent. In a wide sense, phenomenal consciousness is relational if phenomenality is a property a *p*-conscious state has depending on our description or concept of that state (or other, related states that form the context of the *p*-conscious state; Kurthen et al. 1998). Phenomenal consciousness, the being of which uniquely merges with its appearing as far as its like-to-be-ness, or phenomenality, as such is concerned, can then be interpreted as cognitively relational, that is, as a certain way some mental entities are taken by the cognitive system as a whole. Since what something is taken to be depends on a whole cultural background of language, history, scientific knowledge, and so on, phenomenal consciousness can be deconstructed as a cultural byproduct in a wide sense (see Kurthen et al. 1998 for details).

The vehicle theory and the deconstruction of PC. One might then just lose interest in an explanation of phenomenal consciousness, which appears as an ephemeral, description-dependent phenomenon (Dennett [1991] tries to make us lose interest in a similar way). But if phenomenal consciousness itself (and not just our concept of it) can change its character with further cultural and social development, then science – and especially cognitive science – may contribute to this change in a way that helps to make the EG problem disappear. There are at least two ways in which neuroscience can promote a change of the features of conscious states. First, as a theory of the neural correlate of phenomenal consciousness, it can illustrate how the features of phenomenal consciousness will change according to a change in the neural correlate. Second, as a theory that emphasizes convincingly the nonphenomenal aspects of our cognitive lives, it represents one of the influences that may finally make us take our previously phenomenal (or should one say "seemingly phenomenal"?) states as something else, for example, as certain action-related states.

O&O's vehicle theory can contribute to this development, although the choice of the connectionist framework seems to be more important than the commitment to a vehicle theory. Although they thoroughly elaborate on the explanatory power of the vehicle theory with respect to actual features of phenomenal consciousness (see sect. 5), their theory can just as well serve to illustrate the cognitive relationality of phenomenally conscious states – all the more since the mere vehicle-like description of the correlate will not close the explanatory gap (see above). Although the vehicle theory preserves the notion of modularity, it unproblematically allows that the activation patterns assumed to be identical with phenomenally conscious states are permeable to inputs from other distributed patterns that may represent the "taking" of this phenomenally given. Furthermore, the vehicle theory adds an important influence on the phenomenal consciousness patterns that may also alter their phenomenal properties: namely, the "unconscious information" encoded in potentially explicit or tacit representations (sect. 4.2). These nonexplicit representations can encode aspects of a cognitive system like habits, world knowledge, internalized social rules, and so on – aspects that are also dynamic and may influence the phenomenal features of conscious states.

In conclusion, O'Brien & Opie present a connectionist theory of the neural correlate of consciousness, but this theory has no direct impact on the hard problem in the theory of phenomenal consciousness. It can, however, promote a deconstructive theory of it, which may be more promising than a conventional naturalistic theory.

Consciousness should not mean, but be

Dan Lloyd

Department of Philosophy, Trinity College, Hartford, CT 06106.
dan.lloyd@trincoll.edu

Abstract: O'Brien & Opie's vehicle hypothesis is an attractive framework for the study of consciousness. To fully embrace the hypothesis, however, two of the authors' claims should be extended: first, since phenomenal content is entirely dependent on occurrent brain events and only contingently correlated with external events, it is no longer necessary to regard states of consciousness as representations. Second, the authors' insistence that only stable states of a neural network are conscious seems ad hoc.

A certain wise person – Jerry Fodor – once observed that "the form of a philosophical theory, often enough, is: Let's try looking over here" (Fodor 1981, p. 31). Likewise for theories in cognitive science generally, as exemplified by O'Brien & Opie (O&O). They suggest that we try looking for consciousness over here, where here is in the neighborhood of "vehicles" rather than processes. In my opinion, it is the right place to look (see, for example, Lloyd 1989, Ch. 7; 1991; 1992; 1994; 1995a; 1995b; 1996; 1998). O&O motivate the new look with a thorough discussion leading to the conclusion that "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 is (at least) three claims in one: first, the basic claim of vehicularity, namely, that states of consciousness are identical with states individuated by their intrinsic properties, rather than by their functional context. Second, that PDP shows how states of the brain might be complicated enough, and in the right way, to capture the complexity of the phenomenal world. Third, that not every PDP representation is a state of consciousness: only "stable" states will do. I will offer brief friendly amendments to each of these claims.

1. A critique of pure vehicularity. O&O seem to offer the vehicle hypothesis as a candidate scientific theory of consciousness, that is, a contingent framework of sufficient power to account for the phenomena of phenomena. However, their hypothesis may rest on stronger philosophical grounds. The stronger argument, in outline, begins with the tautologies that phenomenal content is experienced, not (merely) ascribed, and that experiences are events or episodes that actually happen – they occur right here (in my brain) and right now. As the phenomenologists might say, experiences are constituted by what occurs for me or to me right now. As such, they are not constituted by, or identified with, anything that is not happening here and now in the brain. That rather obvious statement entails that experiences cannot be identified through any relations they may satisfy. All such relations are accidents with respect to identifying phenomenal content.

The same point is suggested by the good old brain in a vat. Suppose the usual: your brain afloat in broth, with fake inputs and outputs simulating a world. As if that were not implausible enough, now also suppose that the vat scientists did not steal your biological brain, but made it ("you") out of whole neural cloth. And they made it less than a second ago, making sure that it was teeming with all the thoughts that would normally require decades to marinate – they simply installed everything from high school memories to life aspirations to pet peeves. And they are going to shut you down before you get to the end of this sen

The scientists who created and extinguished your flashbulb existence have allowed you a few milliseconds of sentience, an awareness too ephemeral to follow from the usual antecedent causal conditions, and without the usual consequences. But the standard vat-brain intuitions still hold. For that brief second, you enjoyed full blown consciousness, just the same phenomenal state as during a parallel time slice from your mundane unvatted conscious life. The extracranial meddling has no effect on the phenomenal. The thought experiment leads to the same conclusion as the preceding argument: consciousness is what it is, and is not constituted by any relations.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

As an important corollary, then, phenomenal content must be dissociated from representational content. (For a more thorough discussion, see Lloyd 1997.) The vathead just described is an entity with mental states that only seem to be about high school, life aspirations, or pet peeves. With no possible connection to any external reality, its inner states cannot carry information about an outside world, and so cannot represent an outside world (at least by most accounts of representation). O&O are right to discount process theories of the phenomenal, since the process cannot constitute the phenomenal here and now. But they should amend the genus of the phenomenal. It is not a kind of "explicit representation of information," for the same reasons that it is not functionally defined. The phenomenal is simply a complex state. What sort of complexity? Why, the complexity observed in phenomena, of course. Which brings us to:

2. The phenomenal is the neural. The vehicle hypothesis reconstrues phenomenology as one description of the vehicle of consciousness. But that vehicle is also a brain (in humans, at least), so neuroscience ultimately offers the same vehicle under another, different description. This entails that differences, distinctions, categories, and kinds of phenomena correspond to differences, distinctions, categories, and kinds in the brain; the world of our experience really just is an image of the brain at work. From this core identity a vast research program unfolds, which is glimpsed in O&O's discussion of abstract experience (as in the paper they mention, Lloyd 1996). It is probable that the PDP framework by itself is somewhat too general to capture what is distinctive in the phenomenal. After all, some connectionist models employ localist representations, which seem too simple to model the phenomenal. But a future "neurophenomenology" will find within connectionism the specific differentia of consciousness, simultaneously discovering both their phenomenal and their neural descriptors. (Looking here does not entail the miasma of introspectionism, however. Nowadays, for hard phenomenal data one can start with protocol analysis [Ericsson & Simon 1993], or the long-standing results of sensory psychophysics [Clark 1993b].) Meanwhile, regarding the O&O differentia of consciousness, we turn to:

3. Flights and perchings. On "the wonderful stream of consciousness," James commented, "Like a bird's life, it seems to be made of an alternation of flights and perchings" (James 1890, p. 243). For O&O it is perchings only; only stable states of neural networks support the phenomenal. Specifying what is just stable enough to be conscious will be a nasty business, and I suspect any specification will be vulnerable to counterexamples. For example, would a slowly mutating state of the brain not be conscious? From moment to moment it is unstable, but its overall heading is steady enough. Perhaps more important, the insistence on stable states reintroduces a seemingly arbitrary distinction between conscious and unconscious content-bearers. Just such a mysterious difference makes the classical picture deeply unsatisfying, and part of the appeal of the vehicle hypothesis is its complete identification of vehicles with states of consciousness. Moreover, one need not exclude the flights from consciousness. Their flightiness alone accounts for their unmemorability and nonreportability. Why not just have it James's way? Let the stream be undivided.

4. Ars Neurophenomenologica. The Modernist poet Archibald MacLeish prescribed the ideal twentieth century poem in his Ars Poetica (1952), in which the formal sonorities of poetry are celebrated at the expense of meaning and representation ("a poem should not mean, but be"). O&O bring the richness of phenomenology to bear on the richness of cognitive neuroscience, and this too entails a turn away from representation. So, with apologies to MacLeish's ghost, I close by substituting "brain" for "poem" in his manifesto, adapting it for twenty-first century cognitive neurophenomenology.

Ars Neurophenomenologica, after MacLeish's Ars Poetica

A brain should be palpable and mute
As a globed fruit,

Dumb
As old medallions to the thumb,

Silent as the sleeve-worn stone
Of casement ledges where the moss has grown –

A brain should be wordless
As the flight of birds.

A brain should be motionless in time
As the moon climbs,

Leaving, as the moon releases
Twig by twig the night-entangled trees,

Leaving, as the moon behind the winter leaves,
Memory by memory the mind –

A brain should be motionless in time
As the moon climbs.

A brain should be equal to:
Not true.

For all the history of grief
An empty doorway and a maple leaf.

For love
The leaning grasses and two lights above the sea –

A brain should not mean
But be.

Information and appearance

Eoghan Mac Aogáin
Linguistics Institute of Ireland, Dublin 2, Ireland. **eoghan@ite.ie** **www.ite.ie**

Abstract: O'Brien & Opie's connectionist interpretation of "vehicle," "process," and "explicit representation" depends heavily on the notions of "information" and "information processing" that underlie the classic account. When the "cognitivist" assumptions, shared by both accounts, are removed, the connectionist versus classic contrast appears to be between behavioral and linguistic accounts.

I had three major difficulties with O'Brien & Opie's (O&O's) target article.

(1) The distinction between "vehicle" and "process" used to separate classic and connectionist accounts is problematic. O&O explain it as the difference between the things that representations *are* and what they *do*. But even on their own definition of the "process" view it is the content of tokened symbols that corresponds to phenomenal experience, while conversely, however much we try to think of stable activation patterns as free-standing vehicles, there must be a process running in the background to make them so.

(2) The phenomenal world, proposed as the test-bed for the connectionist and classic accounts, seems too loosely defined to give decisive results. It ranges all the way from qualia, to the appearance of things, to the interpretation of pictures, to self-consciousness, to "access" consciousness, and then, by prefixings of "what it is like to," all the way to cognition in general, including language understanding.

(3) The connectionist and classic accounts, as presented by O&O, have too much in common.

I will take the last difficulty first. Like the classicists, O&O use the terms "information" and "information processing" extensively and without comment. I take this to be an indicator of the "cognitivism" that Keijzer and Bem (1996) and others speak about, namely, a tendency to think of all human competences as forms of cognition, and to multiply varieties of representation in an attempt

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

to keep the metaphor going. With it comes the idea of consciousness as an "information-bearing medium."

Here is an alternative. Consciousness is a delayed response to perceptual input. It is the capacity, realised in advanced perception, to block informational and other attitude-forming responses to the perceptual object, leaving us with appearances only. We do it effortlessly in cognition, when we consider the meaning of propositions in isolation from their credibility.

The distinction is also present in perception. It is an "abstract" or "high-order" phenomenon in O&O's sense, as striking as the imposition of perspective on the inverted stairs. We can choose, at will, to see the colour and shape of things only, momentarily filtering out their informational and motivational loadings. We can even consider the possibility of illusion or deception, or wonder whether we are about to awake from a vivid dream. We can "see-as," going directly against belief-compelling forces – the basis for our capacity to see pictures and understand language. Functionally, attitudinal filtering allows beliefs and goals to be formed with greater independence from current perceptual input. It is very likely that the capacity is possessed to some degree by all species that exhibit exploratory behavior.

Information, on the other hand, informs. It causes an increment of credibility, and the representation it achieves has no existence outside this process. On this interpretation of the old signal/symbol, taken from Rozeboom (1972, p. 45), both O&O's account and the classic account, often called "symbolic," fail to produce symbols in the new sense. They cannot block attitude-fixing processes or create mere appearances.

Cognitivists respond by talking about information that is available, explicit, stable, and so on. But unless these terms refer to a separation of content-preserving and attitude-determining processes, they will not work. If they do, then we are left with one form of representation only: explicit (cf Lloyd 1992). Attitude-determining processes do not represent anything. They bind explicit representations into behavioural dispositions.

O&O gather an impressive array of findings and hypotheses around their notion of stable activation, but I would have wished the connectionist perspective to be more evident. Its strength is comprehensiveness. Connectionism links perception, cognition, and behaviour in a single sweep, using a model of broadly based neural responses tuned to the relevant invariances in the environment. Pavlov, the learning theorists, and the early ethologists are its counterparts in the past. The "classic" account, by comparison, offers only a model of a hypothetical internal language.

Smolensky (1988b, p. 63) feared that connectionism, without the behavioural perspective, would become encapsulated inside spurious "levels" of cognition. For example, in the connectionist account facial recognition is a statistical model of inductive behavior in an important human environment. It provides the syntax and semantics. It models the genesis of conviction, in both acquisition and fluent performance. It provides a physiology and predicts behavior. In the cognitivist perspective it becomes an "association" or a "concept," that is, just one more internal word for future use in the language of thought.

O&O started with four levels of representation and ended with two. I think they still have one too many. The simpler the notion of representation, the stronger the link to the environment. As for the phenomenal world, if O&O had taken it in its traditional sense, the world of appearances, as studied by the phenomenologists, psychophysicists, and perceptual psychologists, or "sensory consciousness" as some call it, I think they would have a more favourable setting in which to explore the central paradigm of "stable states in a sea of unconscious causal activity." Colours, shapes, shifting appearances, global effects, and emergent constancies provide the strongest argument for the connectionist account. At any rate I found the examples from this domain the most convincing.

What's new here?

Bruce Mangan

Institute of Cognitive Studies, University of California, Berkeley, Berkeley, CA 94720-3020. **mangan@cogsci.berkeley.edu**

Abstract: O'Brien & Opie's (O&O's) theory demands a view of unconscious processing that is incompatible with virtually all current PDP models of neural activity. Relative to the alternatives, the theory is closer to an AI than a parallel distributed processing (PDP) perspective, and its treatment of phenomenology is ad hoc. It raises at least one important question: Could features of network relaxation be the "switch" that turns an unconscious into a conscious network?

Having argued for some years (1) that phenomena like blind sight and implicit learning are often mediated by nonsensory experiences, (2) that PDP networks naturally model consciousness better than do classical AI models, and (3) that consciousness is an information bearing medium (i.e., roughly, a vehicle) – I am grateful to O'Brien & Opie (O&O) for introducing these possibilities to a wider audience. (See Mangan 1991; 1993b; 1993c for points 1 and 2 above; see Mangan 1993a; 1998 for point 3 above.) And though I cannot speak for him, many of these issues have also been addressed, independently and in somewhat different ways, by Dan Lloyd (1988; 1991; 1995a; 1996).

The new element in O&O's target article is their claim that *only* fully settled networks are represented in consciousness, and that *no* full network relaxation ever takes place unconsciously. This certainly gives them a unique thesis. But it also brings with it a huge problem – for it goes completely against current PDP thinking about the operation of unconscious neural networks in the brain. PDP networks yield the most information when they are fully relaxed; unstable networks are almost always less informative than stable networks.

To accept O&O's theory, we must first conclude that most neural networks forego, for some reason, the very great information processing advantage that full stabilization would otherwise provide them. Why should this shortfall occur? So far as I can see, O&O have neither neural nor behavioral evidence nor PDP theory to justify this extraordinary assertion about the nature of unconscious processing. There is, of course, always the possibility that they are right, and that PDP thinking has made a mistake about one of its fundamental applications. But some sort of strong argument supporting this claim is required, and none is provided. Yet their entire theory of consciousness rises or falls on this point.

In one sense, O&O's concern with phenomenology is on firmer ground. They have a general interest in the "parallel" aspects of experience, which is natural for any PDP attempt to understand consciousness. But their specific task is to show that they can account for our phenomenology on the assumption that only fully relaxed networks occur in consciousness. This presents various obstacles that the other PDP theories of consciousness do not have to face, because they do not make such restrictive assumptions. So O&O have ended up with a fairly rigid theory of conscious, one that distinguishes conscious from unconscious processes more in the mood of classical AI than PDP. As they themselves note: "As was the case with classicism, such a connectionist vehicle theory [i.e., their theory] would embrace the distinction between explicit representation and potentially explicit/tacit representation, as the boundary between the conscious and the unconscious" (sect. 4.2, para. 1).

In consequence, O&O's position is somewhat retrograde. Their theory will not let them take advantage of one of the central ideas found in all other PDP theories of consciousness: that is, that network relaxation can be represented in consciousness even when full stabilization has not occurred, thereby allowing consciousness to represent dynamic as well as fully stable states of neural activity. In rejecting this view, O&O have had to make many ad hoc assumptions about the relation of the brain to conscious experience (see sect. 5) before being able to move on to explain the same sort

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

of "parallel" phenomenology the other theories can address much more directly.

O&O's theory requires that we accept a new view of unconscious processing that is maintained without benefit of supporting evidence, and is at odds with virtually all existing PDP models of neural activity. Even if we set this problem aside, the ability of the theory to handle the phenomenological facts is more cumbersome and ad hoc than the alternatives already developed.

But I suspect all current PDP theories of consciousness will look very deeply flawed in ten or twenty years. The late philosopher of science Paul Feyerabend held that a theory in science can serve many functions besides proclaiming an eternal "truth." Whatever its status, O&O's theory still raises an important question: Is there something about the way a network stabilizes that could "switch" it from unconscious to conscious processing? The standard view is to suspect that activation levels may perform this function. But O&O in effect suggest that the switching mechanism could derive from a more complex aspect of network behavior. I doubt that full network stabilization could work as a consciousness switch, but the general question O&O raise is important, even if no answer is in sight.

A vehicle with no wheels

Drew McDermott

Department of Computer Science, Yale University, New Haven, CT 06520-8285. **drew.mcdermott@yale.edu**

Abstract: O'Brien & Opie's theory fails to address the issue of consciousness and introspection. They take for granted that once something is experienced, it can be commented on. But introspection requires neural structures that, according to their theory, have nothing to do with experience as such. That makes the tight coupling between the two in humans a mystery.

O'Brien & Opie's (O&O's) theory does not really explain anything worth explaining. For example, in section 5.4 they talk about how their theory would explain the ability of people to make fine comparisons among colors. The reason is that in their theory (as in most connectionist theories), everything is represented as a vector in a space and such vector spaces have natural metrics. But where does that get us? On the one hand, *any* reasonable theory will suppose that the space of color representations is as complex as a vector space, so connectionism is not necessary; on the other hand, it is easy to build nonconscious networks of computers, each simulating a neuron, that classify colors as points in a vector space, and so connectionism is not sufficient.

Perhaps I am wrong about this last claim. It is not clear from the article whether O&O believe a network of silicon neurons would be conscious. They clearly do not believe (sect. 5.1) that a network simulated on a single computer would be conscious. They do believe (sect. 4.1) that it is not necessary for consciousness that the neurons be as complex as biological neurons. But they do not quite say whether a network of digital computers, each simulating a neuron of the classic linear-weighted-sigmoid-output variety, would be conscious. I suspect their intuitions are inconsistent on this issue.

The target article is vague on this crucial question, but it has an even bigger bug. It does not really tackle the main problem of phenomenal consciousness for a psychologist, which is to explain why brains think they have phenomenal consciousness. This may sound odd, but I think it is a real problem even for people, like me, who take phenomenal consciousness seriously. Suppose we wonder whether chickens have phenomenal consciousness, as people often do wonder. Whatever the answer, it seems clear that the chicken has considerably less ability to ask the question about itself than people do about themselves. One might speculate that the reason is that we have a "higher" consciousness, but of course the real reason is that we have brains with more complex models of themselves. Some perceptual states, in addition to being links in the sort of causal chain chickens have, are also links in a different kind of causal chain, in which the perceptual states themselves are the object of perception. A person can look for a redder apple, and also notice that one of the apples only appeared to be redder.

O&O do not discuss any of this. They would say, I guess, that the reason for the omission is that we must first figure out what phenomenal consciousness *is*, then worry about how brains react to it. But if experience is just stable activation patterns in neural networks, then there is no obvious way in the authors' framework to distinguish between the ordinary use of a stable activation pattern and its use in reporting an experience.

Furthermore, if phenomenal consciousness is stable activation, then there is no necessary link between consciousness and introspection. Introspection is mainly behavioral, and as such must eventually be mediated by observable events. If I have the belief that I saw red, and utter the words, "I thought I saw red," I do so because of neural events, which in O&O's account occur after the stable activation pattern, and are affected by it but do not affect it. Subtract the report and the experience remains. On this theory there is no reason why conscious experiences should be more accessible to introspection in humans than they are in chickens, and in fact it seems to allow for the bizarre possibility that most of our conscious experiences are not accessible to introspection.

Process theories of consciousness assume that a major problem of consciousness is explaining introspection. It is widely supposed, and O&O agree, that the weakness of these theories is that they do not explain phenomenal consciousness. A process theory explains phenomenal experience, or qualia, by explaining them away, that is, by explaining why a certain sort of self-modeling system would *believe* its experiences had qualia. Such an explanation is not very satisfying, but it may be correct nonetheless, by which I mean that scientifically it may do everything we want. For instance, it might explain everything there is to explain about the circumstances under which brains report qualia; it might work just fine for medical treatment of various brain conditions, and so forth. In other words, were it not for first-person observation, we would never question it.

What I want to call attention to is the peculiar position we will be in if such a process theory gains acceptance as the Third-Person Theory of Consciousness, the one we use in deciding how to do brain surgery, and a theory like O&O's gains acceptance as the First-Person Theory, the one that explains what consciousness really is. The process theory will never actually have to appeal to the vehicle theory for any explanatory power, and vice versa. This outcome would be unsatisfactory. It would be as if there were a theory of "real" economic value and a theory of "apparent" economic value, such that everything to be explained fell into the realm of the "apparent," except what was really valuable. To put the point methodologically: the problem for *any* vehicle theory is that it does not defeat a competing process theory. They are not fighting on the same battlefield.

I think vehicle theories are going to remain nonstarters in the computationalist fleet. Computationalism demands computational links from one event to the next. The medium – that is, the vehicle – of a computational event is by definition invisible to that kind of causal chain. To the extent that the mere vehicle of such an event is important, we have exited off the computational freeway and onto the dirt road of dualism.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

What about the unconscious?

Chris Mortensen

Department of Philosophy, The University of Adelaide, North Terrace, SA 5005 Australia. **cmortens@arts.adelaide.edu.au**

Abstract: O'Brien & Opie do not address the question of the psychotherapeutic role of unconscious representational states such as beliefs. A dilemma is proposed: if they accept the legitimacy of such states then they should modify what they say about dissociation, and if they do not, they owe us an account of why.

O'Brien & Opie (O&O) offer an account of phenomenal consciousness as explicit representations in PDP networks. Explicit representations come down to stable activation patterns. This proposes two distinctions: first, between activation patterns and connection weights; and second, between stable states and transitory states. It is instructive to trace these distinctions via the role of the unconscious in psychopathology, which O&O do not treat but which raises the significant issue of the nature of unconscious beliefs.

I have been struck by the number of practising psychiatrists who find it explanatorily and therapeutically useful to have a notion of the Unconscious. This is hardly to say that the profession as a whole has a commitment to Freudian psychoanalytic theory. However, a minimalist notion of the Unconscious relevant to psychopathology would at least differentiate it from the mere failure to be conscious of our insides. It must imply that the Unconscious contains *mental-like* states such as beliefs. Such beliefs will undoubtedly be representational, just as conscious beliefs are. They can involve heavily symbolic associations, and they are certainly proposed as causally efficacious in behaviour. Not only is the Unconscious unconscious, but it can be difficult to bring its contents to consciousness because of the mechanism of repression; and *part* of the method of much therapy for psychopathology is to identify unconscious beliefs to consciousness, as a necessary step to working against their destructive potential. That is to say, the Unconscious contains stable, causally relevant representational states. What makes them unconscious then? For O&O there can be unconscious representational states; but being unconscious they are not *activation patterns,* but rather *distributions of connection weights,* which are *dispositions.*

Now we can ask: if there can be unconscious beliefs, then what is the difference between one of them and the very same belief when it is conscious, say, if it has been brought to consciousness by a skilful therapist? It would be strange to say that the unconscious belief is a distribution of connection weights, while *the very same* belief when it is conscious is an activation pattern. This would fail to account for what is the same in the situation. O&O must say that the belief in *each* case is the weights/dispositions; while the consciousness of the belief is an activation pattern which can be, but need not be present when the belief is. In a slogan, conscious beliefs are consciousness plus beliefs.

We can ask in what O&O's first distinction, between activation patterns and weights or dispositions, consists. A useful analogy is with the concept of electrical resistance. If we think of synaptic resistances as part of the story of connection weights, then we can say that a simple model for the distinction is the distinction between the presence of a resistor in a circuit and the passage of current through that circuit. The presence of resistance has no effect on the output of a circuit if there is no potential difference between ends of the resistor. Thus, resistance is a *disposition* of a resistor, manifesting itself as having various outputs for various inputs. Vary the resistance, and there will be different outputs for the same inputs.

If electrical resistance provides a model for O&O's first distinction, electrical capacitance provides a model for their second distinction. If we ask how long an activation pattern has to last to qualify as *stable,* their view is that it lasts long enough for its module to produce outputs. This is a *capacitance effect:* the effect of a capacitor is to build up charge over time until a threshold is reached when the capacitor discharges. Unstable activation patterns are *transient,* that is, are changing too quickly to be broadcast generally as the output of the module.

In sum, O&O's theory rests on a pair of legitimate distinctions. Moreover, the presence of both resistance and capacitance effects together gives the behaviour of an oscillator or clock. Thus we can say that O&O's theory has the *consequence* that conscious beings have a sense of *time.*

Even so, while O&O's distinctions are legitimate, one might ask whether they cut nature exactly at the joints. One case is the rapid production of reliable visual memory. Having met a person for a short time only, we can often enough instantly recognise their face years later. If the long-term unactivated memory is laid down quickly, it suggests that the mechanism is one which is very rapidly and sensitively responding to subtle changes in activation patterns. This would seem to be a change less like modification of connection weights and more like modification of computer settings, which are still appropriately dispositional. But this tends to erode the activation/settings distinction, because computer settings are surely stable patterns of charge throughout the circuit. Settings and inputs would thus form hierarchies, which suggests relativising the activation/setting distinction.

An important class of cases are those arising from dissociation, for example, dichotic listening and other manifestations of subliminality. O&O seem fixed on saying that the masked beliefs generated are really conscious, despite their unavailability to verbal report. It would seem that they have to say that even though we are conscious of the stimulus, we do not know that we are conscious. But it is surely a simpler solution to say that we are not conscious of that which we are unable to report. If we allow that the conscious/unconscious distinction is the active/dispositional distinction, then surely there is no problem of identifying masked inputs as modifying connection weights unconsciously; unless one thought that the only way to change connection weights is by subjecting them to *conscious* experience.

We can sum up in the form of a dilemma. If O&O accept the legitimacy of unconscious beliefs, then it is unclear that they have to say what they do say about dissociation. If they do not accept it, then they owe us an account of why therapists have found the unconscious theoretically useful, an account which they have not addressed.

Arguing about consciousness: A blind alley and a red herring

Natika Newton

Philosophy Department, Suffolk County College, Selden, NY 11784. **nnewton@suffolk.lib.ny.us**

Abstract: O'Brien & Opie hold that phenomenal experience should be identified with "stable patterns of activation" across the brain's neural networks, and that this proposal has the potential for closing the 'explanatory gap' between mental states and brain processes. I argue that they have too much respect for the conceivability argument and that their proposal already does much to close the explanatory gap, but that a "perspicuous nexus" can in principle never be achieved.

The blind alley. The conceivability argument assumes that if we can conceive of a state of affairs S, then S is not logically impossible. Applied to mind-brain reductionism, it is used to argue that since we can conceive of any brain state existing without a conscious state existing, conscious states cannot be identical with brain states. It is dismaying that this Cartesian argument is still taken seriously. Considered in one sense it is tautologous; considered in another, it is groundless or incoherent. At best, it has heuristic value for purposes of creative thinking, but it can carry no weight in arguments concerning ontology or metaphysics.[1]

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

In one sense, "conceivable," means "logically possible." To assert that S is conceivable in that sense is simply to *assert* that it is logically possible, that is, that it does not entail a contradiction; it is not to argue that S is logically possible. If it *is* conceivable that a given brain state could occur without a conscious state, then they cannot be identical, because if they were, by virtue of the meaning of "identity," they must coexist. If the issue is the identity of mental and brain states, then asserting conceivability is simply announcing one's position in the debate.

In a looser sense, "conceivable" means "imaginable." Here the appeal is not to the logic of concepts, but to a subjectively generated image. Suppose I wonder if I could conceivably carry my piano to the basement without dropping it. I might try to imagine carrying the piano by generating relevant sensory, motor, and proprioceptive images and the imagery would convince me that I would not be capable of the task.

In this case, I could be right, because the imagery, coming from actual past experiences, is a reliable predictor. But if I were asked whether, in this sense of "conceivable," I could conceive of a brain process without a conscious state, I would be at a loss. First, *can* one imagine a brain process in the sense of forming mental imagery of it? The innumerable microscopic events constituting an active brain process have never been observed, visually or otherwise, by anyone. Second, it is unclear what it is to imagine a brain process occurring *without* a corresponding conscious experience. Would one imagine observing *only* the brain process? One would not observe another's conscious experience whether or not it existed.

It might be objected that one could imagine observing, for example, one's own active pain-sensing mechanisms, while at the same time not feeling pain. But what determines that the brain process I am imagining observing is my *own?* It is unclear what evidence could show that I have indeed imagined observing my own brain process, rather than another's. It is easy to tell a coherent story about observing a brain process that I mistakenly take to be my own. As Williams (1973) has clearly shown, conceivability in this sense is useless for supporting modal claims about mind-body relations.

The red herring. O'Brien & Opie state that their theory has "the potential to close the explanatory gap." I believe that things are more complicated. In one sense they can make a stronger claim; in another, the task is hopeless.

One "mystery" of phenomenal consciousness is its "reflexive" quality: we appear to experience, not only sensory input, but also our own experiencing of the input. A stable pattern of activation allows an extended temporal "present" (James's "specious present"), prolonging it enough for our own responses to be experienced along with the continuing stream of input. Proprioceptive data from responses to immediately past input, held in working memory, is blended with new input, yielding an experience of ourselves experiencing the input. The blend is unified in the way explained by the authors. This phenomenon adds temporal depth to the experienced present moment, and explains the almost paradoxical "reflexive" aspect of consciousness (Ellis & Newton 1998; Newton 1996). To illustrate: consider a being whose brain works like ours except with no ability to retrieve past responses to sensory and proprioceptive input. We might grant that at each instant or time-slice, such a being possesses all the levels of *information* that we derive *from* conscious experience about internal and external events. But it would not have *phenomenal experience,* having no way to compare the current events with any others, and hence no way to appreciate what they "are like" (or "*not* like"). Our responses to stimuli, in the form of just-past representations superposed with new input, linger; we can thus compare and evaluate them (Damasio 1994). That is why they are "like" something.

A fully "perspicuous nexus" between mechanism and conscious experience, however, may be impossible. A subject perceives its world *as* the intentional object of an embodied conscious agency, a self. Scientific observations of an active, stimulated brain, on the other hand, are observations of an external, nonintentional mechanism (the self of the observing scientist is not the self of the observed brain). Passing from one "view" to the other would entail an aspect (Gestalt) shift. As with the famous duck/rabbit drawing, I can take the "intentional stance" toward my own perception of the world, or the "machine" stance toward my brain. Both stances are made possible by my physical components. But no single coherent (nondisjunctive) description will capture the aspects of *both* stances, any more than such a description will capture both the duckness and the rabbitness of the single drawing. The viewer, and the scientist *qua* conscious subject, must choose or alternate between them.

NOTE

1. For a related discussion of the argument see Tye 1983; also Newton 1989.

Why information?

Joseph O'Rourke

Department of Computer Science, Smith College, Northampton, MA 01063.
orourke@cs.smith.edu cs.smith.edu/~orourke/

Abstract: O'Brien & Opie's admirably sharp hypothesis gains some of its force by ignoring distinctions in murky areas. I attempt to agitate the waters by suggesting that process and vehicle theories are not so different, that classicism can support a vehicle theory, and that several of the key concepts underlying their theory are less clear than depicted. The connection to information I find especially tenuous. Finally, I address the implications of their theory for unconscious thought.

1. Attention. When attending to the work in front of me, I am not conscious of "the chair pressing against [my] body," although if asked to turn the "searchlight" of my attention to this pressure, I do notice it. O'Brien & Opie (O&O) include the chair-pressing as part of the phenomenal field of experience, much as Block considers unnoticed background jackhammer noise part of P-consciousness (phenomenal consciousness) (Block 1995, p. 234). But many process theories of consciousness, such as Baars's, focus more on "events in the bright spot on stage" (of the "theatre of consciousness"), for only these are "strictly conscious" (Baars 1997, p. 303). As O&O themselves suggest "more intense processing" to account for attention, it may be that the divergence between process and vehicle theories of consciousness is partly due to the different senses of "consciousness" used by their proponents.

2. A defense of classicism. The operation of a digital computer is determined by its architecture, the program, and the data. There are no clear lines of distinction separating these three: the program can be reduced by complicating the hardware or by swelling the data. Nevertheless, a classicist who desired a vehicle theory could draw the lines and identify the phenomenal field with representing data. Learning need not be "merely a process that reconfigures the brain's functional architecture" – it reconfigures the program, precisely how many classical machine learning programs operate.

A symbol in an executing classical program is a stable pattern of (perhaps thousands of) bits distributed in memory chips, explicitly representing information. O&O claim that "the stable activation pattern is absent in digital simulations of PDP systems," but the basis of cache memory is exploiting the stable activation patterns exhibited by running programs. Here I am capitalizing on the ambiguity of some of the key notions underlying O&O's theory, to which I now turn.

3. Fuzzy concepts. That there is some fuzziness in the concept of "explicit representation" is demonstrated by the number of philosophers who, with Block (1995, p. 279) say that "the phenomenal content of an experience goes beyond its representational content" – in contrast to O&O's commitment to the thesis that "all phenomenal experience is representational." I suggest

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

again that the differences may be more definitional than substantive. The notion of "physically discrete" is perhaps clear enough, but it doesn't distinguish between distributed neural net patterns and widely scattered bits of a classical token. Moreover, the requirement of "causal potency," which is layered on top of the main hypothesis when rejecting classicism, is slippery: Is there a principled way of distinguishing those features that "are actually doing the causing as opposed to figuring in a convenient description of the causal process"? (Garfield 1997, n. 17).

Although O&O make a strong case for the importance of stable activation patterns, it is an assumption of their model that only a "constant rate" of firing "can facilitate meaningful communication between PDP networks." It makes intuitive sense that we can "feel" the coordinated firing of our neurons, but it is equally plausible that we can feel patterns in time: waves (perhaps at the oft-cited 40 Hz) pulsing, or other "complex dynamical features" that neural nets exhibit. Processing itself can be viewed as a (time-varying) pattern: Why should we feel only stationary patterns?

4. Why information? Most significantly, why should patterns encoding *information* be those felt? It seems the only connection to information is the intuition that consciousness is "consciousness 'of something'." Even if we accept that only time-stable patterns have a feel, why only those patterns that explicitly encode information? Presumably what constitutes information is species-relative, and perhaps even observer-relative: Is a pattern that bears information for me a pattern that must bear information for you? (Note the relation here to Searle's insistence [1992, p. 209] that "syntax is essentially an observer-relative notion.") Would O&O claim that every stable activation pattern explicitly encodes information? If so, information is irrelevant to their theory – stable patterns suffice; if not, an account is needed to explain: Why information?

5. Unconscious thought. Perhaps the most jarring implication of O&O's theory is that all unconscious mental activity is free of explicit representation: It is all just "relaxation processes." This inverts Baars's "big puzzle": "Why is the conscious aspect so limited, and the unconscious part so vast?" (Baars 1997, p. 294) Now we must wonder if the unconscious is so limited. Is it plausible that, between time t when I say "I know her name – it will come to me," and t + five minutes later when it suddenly does come to me, no information is explicitly represented in the patterns of neural activity that constitute the unconscious search process? Just barely.

Higher order thinking

Josef Perner[a] and Zoltan Dienes[b]

[a]*Institut fuer Psychologie, Universitaet Salzburg, A-5020 Salzburg, Austria;* [b]*Experimental Psychology, University of Sussex, Brighton, Sussex BN1 9QG, England.* **josef.perner@sbg.ac.at www.sbg.ac.at/psy/staff/perner.htm dienes@epunix.susx.ac.uk**

Abstract: O'Brien & Opie's position is consistent with the existence of implicit learning and subliminal perception below a subjective threshold but it is inconsistent with various other findings in the literature. The main problem with the theory is that it attributes consciousness to too many things. Incorporating the higher order thought theory renders their position more plausible.

O'Brien & Opie (O&O) provide some interesting discussion of the relationship between representation and consciousness. We are not persuaded by their argument, however, for both empirical and theoretical reasons. Empirically, their review of the literature on the dissociation of conscious experience and mental representation is selective and dated. The criticisms of many of the studies have largely already been dealt with. Blindsight is attacked on the stray-light hypothesis, which has been quite extensively countered (e.g., Weiskrantz 1987). There is no discussion of the more recent evidence on the implicit-explicit dissociation in the visual domain (e.g., Milner & Goodale 1995). These are cases where on the vehicle theory behaviour is governed by presumably "explicit representation" that is not accompanied by phenomenal experience (in fact, the phenomenal experience contradicts the behaviour). Similarly, the critique of subliminal perception rehashes old arguments already dealt with by Marcel's (1983) experiments. For example, Marcel (1983, Experiment 4) interspersed threshold determination and priming trials to counteract the effect of any general drift in light adaptation throughout the experiment. He also found equivalent levels of priming for masked and unmasked primes (Experiment 3), which rules out the claim that "the (small) degree of priming that occurs may well be entirely due to chance conscious events." More recent studies (such as that of Neumann and Klotz [1994] finding absolute $d' = 0$ and still there is a clear effect on RT, although with geometric shapes and not word meaning) are not mentioned. In sum, we believe the existing evidence for a dissociation between conscious experience and mental representation is more compelling than that presented by O&O.

As an aside, O&O's position does not rule out implicit learning; if anything, implicit learning is at the core of their theory. A person may be aware of elements of a stimulus and of their behaviour because these are coded by stable activation patterns. However, in many cases, the relationships between the elements and between elements and behaviour will be learned by changing connection weights. Connection weights, by O&O's theory, do not support conscious experience. Hence people will frequently learn of the relationships between stimuli without being aware of those relationships, which is just how O&O define implicit learning (for an argument that this is what actually occurs, see Dienes & Berry 1997). O&O's theory also does not rule out subliminal perception as defined by a "subjective threshold"; that is, visual input may lead to a stable activation pattern and hence some conscious experience, but not the experience of seeing something; so a person can legitimately claim they did not see a word. Subliminal perception in this sense is apparently well accepted by experimental psychologists (Greenwald 1992).

Of course, for any particular piece of empirical evidence for a dissociation between representation and phenomenal experience, there are fresh counterexplanations that can be raised. Maybe the most compelling argument for a dissociation in our minds is a logical one: there is no reason to believe that there should be a necessary or even a strong relationship between the representation of X and consciousness of X. O&O recognize this in allowing unconscious classical representations. Why should a connectionist style of representation be any different? The trouble with their vehicle theory of consciousness is that it would be easy to set up a real PDP network made up of electronic chips with a stable pattern of activation, which we would all agree had no more consciousness than a thermostat. What difference would it make if the chips were replaced with neurons? What if I cut off a bit of network from the brain and maintained its activation electronically? What of a pattern of sustained activation in the spinal cord? The vehicle theory of consciousness simply does not make the link to phenomenal experience clear. Section 5 feeds hopes of a better understanding that will make it inconceivable to think otherwise, but until that stage is reached, the theory does not make the case.

We believe that a (necessary but perhaps not sufficient) missing link is provided by the higher-order-thought Theory of Consciousness (e.g., Carruthers 1996). The basic insight is that to be conscious of some state of affairs (e.g., that the banana in my hand is yellow) I must also be aware of the mental state by which I behold this state of affairs (i.e., that I see that the banana is yellow). There is something intuitively correct about this claim, because it is inconceivable that I could sincerely claim, "I am conscious of this banana being yellow" and at the same time deny having any knowledge about whether I see the banana, or hear about it, or just know of it, or whether it is I who see it, and so forth. That is, it is a necessary condition for consciousness of a fact X that I entertain a higher mental state (second order thought) that represents the first order mental state with the content X.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

A representation of X that does not produce a representation of the propositional attitude by which X is beheld would not be conscious, in this account, thereby contradicting O&O's theory. But a representation (call it Y) of "I am thinking X" does allow consciousness of X. In a way, this is a vehicle theory – the possession of a suitable representation (Y) is the necessary and sufficient condition of consciousness. On the other hand, representations like Y could only emerge because suitable processes operate on representations like X – operations rich enough for us to attribute mental-state terms such as "thinking" to them. Thus, while a free standing three-layer neural network could be conscious in O&O's theory, it would not be on the higher order thought theory, because such a network could not legitimately represent itself as thinking anything. Whatever our ultimate theory of what thinking is, a system would have to approximate the kind of information processing activities that humans get up to before we would be willing to attribute the label "thinking" to it.

Sorites paradox and conscious experience

Tamás Pólya[a] and László Tarnay[b]

[a]*Department of Linguistics;* [b]*Department of Philosophy, Janus Pannonius University, Pécs, Ifjúság út 6, H-7624, Hungary.* **polya@btkstud.jpte.hu**
tarnay@btk.jpte.hu

Abstract: The theory of consciousness proposed by O'Brien & Opie is open to the Sorites paradox, for it defines a consciousness system internally in terms of computationally relevant units which add up to consciousness only if sufficient in number. The Sorites effect applies on the assumed level of features.

There are striking *structural* similarities between O'Brien & Opie's (O&O) account of phenomenal consciousness and Dennett's Multiple Drafts Model, in that phenomenal experience for O&O and operations of thought for Dennett consist of a multitude of distinct *content-fixations* distributed across the brain. Furthermore, while Dennett and Kinsbourne (1992) ask "whether any particular content thus discriminated will eventually appear as an element in conscious experience," O&O define consciousness (explicit information coding) as the generation of a stable activation pattern out of unconscious causal activity in the brain. However, Dennett and Kinsbourne and O&O differ in the level at which they identify consciousness (neurological for O&O, some higher for Dennett & Kinsbourne). Dennett proposes a process theory of consciousness, while O&O offer a vehicle theory. It is precisely O&O's claim that phenomenal experience is identified with "an intrinsic, physical, intra-network property of the brain's neural networks" that exposes their approach to a criticism based on the Sorites paradox.

In the philosophical literature, the Sorites argument is aimed at exposing the indeterminateness of boundaries in applying certain predicates in natural language. Although there is still considerable debate about whether the vagueness of such predicates (recurrent ones are "be a heap" or "bald") is due to blurred boundaries in the world, or inadequate definitional/conceptual criteria, or limits in application, the puzzle remains to be that of identifying in terms of constituent structure.

Taking an example kindly provided by O&O, according to which "consciousness . . . is a rich tapestry woven from many threads," one can wonder whether, if a single thread does not constitute a tapestry, adding a thread, or two, or three, or four, and so on will constitute it, it being obvious that, say, a hundred threads do make up a rich one; or starting at the other end, the puzzle is whether by removing threads one by one from a tapestry of a hundred threads, a precise step can be reached when it ceases to be a tapestry, or by induction, even a tapestry of no threads remains one.

The crucial problem in cognitive research pursued from a *system-internal* point of view resides in determining at exactly which level of structural complexity consciousness "appears." For suppose science produced sophisticated connectionist models, differing only in their structural complexity: Which of them could be called conscious? The number of units, their arrangements, the quality and quantity of their links, and the system's learning properties may all, in principle, contribute to making a system more complex, and, according to a strictly system-internal approach, respectively to making it conscious. The predicate "be conscious" seems to be always Sorites-like to some of the constituents contributing to the appearance of consciousness.

The Sorites argument concerning consciousness in connectionist networks turns on how one defines the units of constituent structure. On O&O's "vehicle" account we are provided with the definition that a stable activation pattern *is* phenomenal experience. Since they claim that instantaneous phenomenal experience is a "complex aggregate state composed of a large number of distinct phenomenal elements," one may wish to level the Sorites argument against their identity relation at the level of how many stable activation patterns make up a conscious experience. O&O may retort that a single stable activation pattern would already *eo ipso* constitute phenomenal experience, while human consciousness is *de facto* always a lump of such states. Even if we are not totally convinced of the viability of the escape route above, let us apply the Sorites to the most distinctive aspect of their definition, the *stability* of states. There are two ways to do it: first, *spatially*, to the magnitude of the physically connected elements, the number of neurons or neuronal connections at a stable state, on which a single phenomenal experience is supervenient. Yet O&O may regulate this to empirical neurological research. But no similar alternative is open to them in the other – *temporal* – issue. O&O say, citing Churchland and Sejnowski (1992), that given their chemical dynamics, "stabilizations can occur very rapidly"; the question now is not how many stable activation patterns brainy networks do actually generate per second, but rather how one can conceive of *stability as such.* That is, what time interval defines the rapid sequence of stable activation patterns? Theoretically put: How should we conceive of rapidity *vis à vis* continual changing in digital simulation? And empirically: How should we check that a given degree of stability in fact appeared?

We cannot rely on phenomenal experience, because according to O&O "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." One can infer from the case analyses referred to and also that phenomenal experience is by definition explicit information coding, that the existence of stable activation patterns is justified by introspection and verbal reports of the subjects under scrutiny. Yet introspection and verbalization are not entirely reliable.

Conversely, how do we know that when verbal reports attest to the presence of phenomenal experience, there is no unstable activation pattern that essentially contributes to explicit information coding? Hence, in order to justify their theory of the correlation, let alone the identity, of stable activation patterns and phenomenal experience, O&O need criteria independent of the terms of their definition. Verbal reports may be unreliable, or lacking altogether (as in animals), which shows O&O's hypothesis untestable.

Getting the vehicle moving

George N. Reeke, Jr.

Laboratory of Biological Modelling, The Rockefeller University, New York, NY 10021. **reeke@lobimo.rockefeller.edu**
www.rockefeller.edu/research/heads.htm

Abstract: O'Brien & Opie present an attractive alternative to the popular but flawed computational process approach to conscious awareness. Their "vehicle" theory, however, is itself seriously flawed by overstrict allegiance to the notion that explicit representation and stability are defining hall-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

marks of consciously experienced neural activity patterns. Including reentrant interactions among time-varying patterns in different brain areas can begin to repair their theory.

While presenting in detail only two of the many possible objections to the classical computational theory of mind – that it does not allow Dennett's (1982) explicit states to be unconscious and that it cannot account for learning – O'Brien & Opie (O&O) make a strong case for an alternative based on representation rather than process. In their version of connectionism, phenomenal experience, in accord with common intuition, is a complex amalgam of patterns shaped by multiple sensory inputs and information processing correctly depends on phenomenal experience rather than the reverse. It is unfortunate, then, that in championing this appealing approach, O&O have rigidly specified what makes patterns of activity eligible for consciousness while almost completely neglecting the connections among them. Connections allow the dynamic interplay of inactivation patterns to form what Edelman (1987) has called "global maps," which allow behavioral implication (loosely "meaning") to become congruent across diverse brain regions without agreed codes, and which allow images and thoughts to flicker continually in and out of conscious awareness. In my opinion, O&O err in focussing exclusively on just two requirements for their "representational vehicles" to be conscious – explicitness and stability.

1. Explicitness. O&O base their analysis on Dennett's (1982) classification of representations of information as being explicit, implicit, potentially explicit, or tacit. The essential kernel of O&O's theory is that phenomenal experience is *identical* to the brain's explicit representation of information (sect. 5, para. 1). Although it is attractively simple, this postulate draws the line in the wrong place, forcing laborious reinterpretations of apparent dissociations of conscious experience and explicit representation in the literature of dichotic listening, blindsight, implicit learning, and so forth. But why, after all, must explicitness be the necessary and sufficient condition for conscious awareness of stable patterns?[1] There are many other ways to classify patterns of neural activity than to equate them to representations and apply Dennett's (1982) formulation to them. Distinctions based on spatial extent and level of activation, location in the brain, connectivity to other areas, concurrent status of nonspecific activating systems, and even neurotransmitter type all come to mind as possible factors that could contribute to conscious awareness. Furthermore, in a nonclassical theory, one would like to treat representation as emerging from neural activity, not as a primary entity. It seems incorrect to base eligibility for conscious awareness on a classification of a secondary, purely computational construct. I therefore suggest that O&O amend their theory to state simply that certain kinds of activation patterns, the detailed characterization of which is left to the future, lead to phenomenal awareness and/or the formation of representations. When this is done, inexplicit subconscious and unconscious influences of all kinds can perform their expected Freudian role, shaping the "view of the road" from the conscious vehicle.

2. Stability. O&O's absolute demand that activation patterns be stable to be conscious leads to additional unnecessary difficulties for their theory: stability, even in suitably quantized chunks of time, is incompatible with the smooth flow of phenomenal experience, which is never stable, and leads them to the absurd claim (sect. 5.1) that only stable patterns can be communicated between networks, when it is basic to information theory that stable signals can carry no information at all.[2] Why do O&O feel obliged to defend stability so strongly? Stability is necessary to their vehicles because, although they have rejected computational process as underlying consciousness, they have not also rejected "encodingism" (Bickhard & Terveen 1995), even though the only purpose of codes is to support computations, making codes unnecessary where there are no computations. Inasmuch as one of the best known methods of instantiating codes in neural networks is via stable patterns of activation (Hopfield 1982), O&O opt for requiring stability.

The simplest way to repair this awkward vehicular engine malfunction is simply to jettison the stability requirement. Allow dynamic patterns to be conscious. With the theory revised in this way, connectivity and the reentrant interaction of patterns can assume their rightful roles as primary elements in the generation of phenomenal experience. Although it might appear that by this device process has been allowed to reenter the scene as cause of phenomenal experience, this is not the case; a stationary vehicle has merely been replaced by a moving one. There is still no need, as in computational process theories, for an observer to interpret patterns of activation or for a processing cycle.

3. Conclusion. Having loosened the requirements for explicitness and stability, we can look again at the question of computational resource. O&O had to fudge their analysis because under reasonable definitions classical and connectionist systems in fact have the same computational capabilities (Lloyd 1996). The problem hinges, as I have pointed out elsewhere (Reeke 1996), on just what definitions are used. Once it is conceded that what is going on is not a classical computation, then differences in computational resources as usually defined are not an issue.

O&O's perception that a serious alternative to the process theory is needed (and possible) follows, I believe, from the fundamental bankruptcy of the process theory. Whether a purely vehicle theory, even one patched up along the lines I have suggested, can do any better remains to be seen. Surely any finally successful theory will contain as prominent elements the ever-changing interplay of patterns of activity in subnetworks specialized in various ways for different modalities and tasks. These subnetworks will be seen to communicate changing, not stable, patterns of activity among themselves via reentrant connection pathways.

NOTES

1. O&O unfairly tar their straw man classical vehicle theory with the same brush, forcing all unconscious representations to be tacit, for example.

2. Naturally, O&O try to escape this problem by invoking a succession of stable states, but it seems better to abandon stability altogether.

What has consciousness to do with explicit representations and stable activation vectors?

Jürgen Schröder

Hanse Institute for Advanced Study, 27749 Delmenhorst, Germany.
jschroel@urz-mail.urz.uni-heidelberg.de

Abstract: To assess O'Brien & Opie's connectionist vehicle theory of consciousness, (1) it is not enough to point to the methodological weakness of certain experiments (dichotic listening, etc.). Successful cognitive theories postulating explicit unconscious representations have to be taken into account as well. (2) The distinction between vehicle and process theories cannot be drawn in the way envisaged by the authors because a representation's explicitness depends not only on its structural but also on its processing properties. (3) The stability of an activation vector is not very suitable for implementing the explicitness of a representation.

1. Arguments for unconscious representations from successful cognitive theories. To argue for their identity thesis, O'Brien & Opie (O&O) need to set to rest suspicion that explicit unconscious representations might be involved in cognitive processes. The most pressing worries, they assume, arise from various strands of psychological and neurological research (dichotic listening, blindsight, and implicit learning), which purports to show that there are unconscious cognitive processes and therefore unconscious (explicit) representations. Another class of arguments which the authors did not address is perhaps as powerful as

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

the direct evidence that derives from the studies mentioned. Every successful theory of a cognitive capacity implies (in a realist conception of science) that the entities postulated by the theory exist. If successful theories of cognitive capacities postulate representations of whose contents we are not aware, then these representations are assumed to exist. One uncontroversial example of such a theory is Marr's (1982) theory of vision. Some levels of representation correspond to our intuitive idea of what we experience perceptually (the 2½-D and the 3-D level), but the image and the primal sketch, representing light intensity and intensity changes, respectively, do not seem to have contents we are aware of. What are we to do with these representations? Should we try to do without them just because our favorite theory of consciousness says there cannot be such things?

2. The concept of explicit representations. Central to O&O's vehicle theory of consciousness is the notion of a representation's explicitness. When is a representation explicit? Kirsh (1991) has argued that our intuitions about explicitness are based on structural criteria (such as a definite location of representations) as well as on processing criteria (such as the direct availability of informational content for a system). However, in some cases the two sets of criteria yield conflicting results. An example would be an encrypted text which counts as an explicit representation according to the structural criteria and as implicit according to the processing criteria. To avoid such conflicts, Kirsh formulated four conditions on explicitness which preserve what is valid in our intuitions and remove the troublemaking aspects (ibid., pp. 350–60). Kirsh's third condition is that a representation is explicit if it is either readable in constant time or sufficiently small to fall within the attention span of an operator (ibid., p. 358). So the processing is included in the resulting set of consistent criteria for explicitness. If an explicit representation is one which is easily processed, that is, whose information is easily recovered and put to work for the task at hand, then being a stable activation vector *per se* does not count as an explicit representation. But if being explicit is, among other things, being easily recoverable, then this is not adequate to distinguish a pure vehicle theory, which would be "pure" by an exclusively structural criterion for explicitness, from a pure process theory. If conscious events are to be identified with the tokenings of explicit representations, the processing aspect must be included and the proposed distinction between vehicle and process theories must be drawn in a different way.

A further problem with the structural criterion of explicitness is that it excludes distributed activation vectors from being explicit. For example, tensor products (Smolensky 1991) representing propositions would not be explicit and therefore would not count as realizers of conscious states. This seems to be a counterintuitive consequence of the structural criterion because it is not clear that our propositional thoughts are *not* represented in a distributed form. According to the structural criterion and the identity hypothesis, however, our conscious thoughts *could not* be represented that way.

3. Why are stable activation vectors necessary? O&O mention two reasons why the stability of activation vectors is crucial for consciousness: one is that the absence of conscious events during dreamless sleep can be explained by the absence of stable activation vectors. The other concerns the interaction of two or more networks. If one network takes the other's output as its input, it can only settle into a stable state if its input is stable. Granting the validity of the first reason (although one may doubt that it is the unstability of activation vectors instead of the absence of a certain kind of processing that really accounts for the absence of consciousness), it seems that the second makes it difficult to explain certain experiences. For example, it is a common observation that in situations where an impending accident can be avoided, one acts *before* one becomes aware of the danger, for example, one hits the brakes, and only then does one consciously see that there is a car coming from the right. According to O&O, however, it should be the other way around.

According to O&O, an activation vector is stable when "its constituent neurons are firing simultaneously at a constant rate" (sect. 5.1). Because this characterization does not make reference to any time scale, a vector could be stable if the simultaneous firing lasted a second or a fraction of a second. Simultaneity alone seems to be too weak a condition for stability, because without relativizing it to a particular time scale every activation vector could be stable; that is, there would be no unstable vectors. This difficulty cannot be avoided by a definition that exploits the relation between networks. If only those output vectors of a network A were stable that yield stable output vectors in a network B, then nothing would be achieved because stability would be defined in terms of stability. In artificial neural networks a stable output vector (or overall state) is that vector which no longer changes when the net goes through further processing cycles, *while the input is constant.* The problem with this criterion, when it is applied to real brains, is that input in real networks is not constant, so a changing output vector could either be the result of a changing input (and then the *previous* output vector would have been a stable one) *or* it could be a transient output vector (and therefore unstable).

Finally it could be said that the coherent interpretation of an activation vector decides on whether or not it is stable. The problem with this suggestion is that it simply does not work. Take any connectionist network that does some classification. Suppose it classifies fruit and has ten possible output classes. Then, if, say, a strawberry representation is activated at the input layer, there might be either transient output vectors representing raspberries or vectors representing fantasy fruit (e.g., half strawberry and half apple). In all these cases, however, the interpretation is coherent even if some of the representations do not represent things that exist.

What unifies experiences generated by different parts of my brain?

Eric Schwitzgebel

Department of Philosophy, University of California, Riverside, CA 92521-0201. **eschwitz@citrus.ucr.edu**

Abstract: Neither of the explanations O'Brien & Opie offer to account for "subject unity" succeeds. Subject unity cannot arise from constructed personal narratives, because such narratives presuppose a prior unity of experience. Subject unity also cannot arise from projection of experiences to the same position in space, as reflection on pregnant women and the spatially deluded reveals.

If consciousness is distributed throughout the brain, as O'Brien & Opie (O&O) contend in section 5.2, the question arises, what makes the conscious experiences generated by two different parts of my brain both part of a single, unified consciousness – *my* consciousness – rather than parts of two separate consciousnesses, as we might suppose a two-headed, two-brained creature to have? Why does the visual cortex not just have visual cortex experiences and the thalamus just have thalamus experiences, each knowable to the other only from the "outside," just as my wife's experiences, even if they are intimately known to me, are not knowable to me in quite the same way my own experiences are? O&O call this question the question of "subject unity," and in section 5.3 they offer two possible answers consistent with their vehicle theory of consciousness. Neither answer is adequate.

O&O suggest, first, that we can treat subject unity "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." They regard this narrative as a "serial stream of self-directed thought" that comes out of the language centers of the brain and thus as something that languageless animals lack.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

This treatment of subject unity is unconvincing for two reasons. First, because languageless animals cannot produce personal linguistic narratives, such narratives cannot be invoked to explain their subject unity. Yet it seems plausible to suppose that there is some sort of subject unity in at least such languageless creatures as infants, dogs, and apes. Some mechanism other than a language-based one, then, must be at work in these creatures. But would not such a mechanism, if it has to exist, also plausibly be the one responsible for subject unity in adult human beings as well? O&O's maneuver seems to commit them to positing two mechanisms to do the work of one.

A more basic objection to O&O's narrative solution to the problem of subject unity is this: the construction of a personal narrative presupposes and depends on the possession of the experiences by a single, unified consciousness and so cannot possibly serve as the causal mechanism that unites those experiences into one consciousness. Unless the experiences are already *my* experiences, I cannot fabricate a narrative of the right sort from them. A one-headed person presumably constructs one narrative rather than two, three, or fifteen hundred. O&O have not explained why this is so, and thus have not yet answered the challenge they have posed themselves of explaining how consciousness arising from different parts of the brain comes together into unified experience.

O&O argue that there is a second way to explain subject unity consistently with their theory of consciousness, "in terms of the confluence of the points of view generated by the individual phenomenal elements that make up our instantaneous conscious experience." These points of view are apparently spatial, since "each of them encompasses a space with a privileged locus, a point with respect to which every content is projected" and so "generate a single phenomenal subject located at a particular point in space." If O&O do not mean to understand "points of view" as literally spatial but rather as a metaphorical way of identifying which phenomenal stream different experiences belong to, then they have begged the question. But neither will a literally spatial construal of "point of view" do the work O&O need it to do. For starters, one might legitimately wonder whether a vague feeling of depression or loneliness has a definite location in subjective space.

Setting that issue aside, however, the question arises how close in space is close enough to guarantee subject unity? A two-headed creature might have two distinct consciousnesses with their loci of subject unity very near each other, even overlapping if the heads both receive sensations from certain parts of their common body (assuming that a sensation in the toe is located by the subject in the toe). At the same time, experiences projected to two different parts of the body or the head of a normal individual must be close enough *not* to count as belonging to the different subjects. Furthermore, consider the case of a woman nine months pregnant. Perhaps the fetus at this point is conscious – I see no reason to rule this out in principle. But if it is, its conscious experiences are not unified subjectively with its mother's, despite their loci being very near and perhaps overlapping in space. The raw distance of two experiences from each other in space cannot be the factor that determines whether those experiences belong to the same subject. Something else must be invoked, but O&O give us no clue what.

Also problematic for a spatial account of subject unity are cases of delusion: Person A mistakenly thinks he is at such-and-such a location where in fact some other Person B is. Their subjective experiences are projected to the same point in real space. Perhaps O&O will say that where the experiences are projected in real space doesn't matter: what counts is where they are projected in "subjective" space. But it would seem that the only way to guarantee that the projections of two different people would not overlap in subjective space would be to treat the subjective spaces of different consciousness as in some way incommensurable – but that would require antecedently determining which consciousness is which, and would thus beg the question.

Perhaps O&O could address the unity of consciousness issue by appealing to relations of informational access that different conscious parts of the brain have to each other. Of course it would have to be explained how this informational access was different in kind from the informational access people have to other people's brain states and why those differences would be sufficient to explain unity of consciousness. I do not see, however, why O&O could not make some more conventional approach like this harmonious with their account.

The slippery slopes of connectionist consciousness

John G. Taylor

Department of Mathematics, King's College, London WC2R2LS, United Kingdom, and Institute of Medicine, Research Centre Juelich, D52425, Germany.

john.g.taylor@kcl.ac.uk taylor@medicom03.ime.kfa-juelich.de
www.mth.kcl.ac.uk/research/staff/jg_taylor.html

Abstract: The basic postulate that consciousness arises from stable states of recurrent activity is shown to need considerable modification from our current knowledge of the neural networks of the brain. Some of these modifications are outlined.

This is a brave and adventurous article by two philosophers who have dared to venture out into the treacherous waters of connectionism and neuroscience to propose what they claim is a fundamentally new postulate for consciousness: "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 postulate is developed somewhat briefly at the end of the target article after a spirited defence of the connectionist approach to the brain has been set against the competing classicist AI view. I do not want to consider that part of the battle being waged, because in any case I am biased as a neural network researcher over the last nearly 30 years. However, I expect neural networks and the classical view to be fused in the end, with the classicist' view providing many useful and important hints to help understand higher cognitive processes. Yet I must take issue with the basic postulate presented in O'Brien & Opie's (O&O's) target article, since I think that (a) it stems from an incomplete understanding of the neural networks of the brain, (b) it is very likely wrong, and (c) it can be replaced by a more realistic set of postulates that take fuller account of the complexities of consciousness (which the authors have ignored completely in setting up their basic postulate).

Let me start with the neural networks of the brain. They do not "relax into stable states," because if they did they would stay in them for ever (assuming "stable" has its usual sense). That is clearly impossible since the brain would soon fill up with unwanted activity. Nor are models of neural networks only of the relaxation sort. These lie at one extreme end of a spectrum in which the other end is occupied by feedforward nets, in which the relaxation is trivial (immediate) because there is no recurrence or feedback of output to keep circulating round as part of the relaxing process.

There is considerable feedforward processing observed in the brain, as well as clear feedback from frontal regions to posterior ones. There are sites where clear continued activity (over 20 or more seconds) is observed in the frontal lobes (Fuster 1989; Goldman-Rakic 1996). Yet this activity involves effortful processing unlike that during more passive conscious experience, as brain imaging now shows: "soft" problems (subspan, needing less than a few seconds worth of previous activity) only activate posterior areas while "hard" ones light up the frontal lobes. These frontal sites are quite different from the posterior region, where no more than about a second's-worth of continued activity is ever observed (Lu et al. 1992). Even then there is considerable uncertainty as to the

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

nature of this continued activity; it is very important to probe this so as to understand working memory (Baddeley 1986), which is suggested as a basic component of consciousness by numerous thinkers (Taylor 1998a).

I have already indicated one way O&O's basic postulate is wrong: neural activity causing consciousness must die away. But there does not seem to be suitable long-term recurrence in a given cortical module to create continued activity except either very locally in cortical sites or in hippocampus. This is a well-known candidate for relaxation due to its abundant lateral connectivity. However, there are patients with no hippocampus but still preserved consciousness (although of a somewhat breathless sort, with no past). So stability brought about by recurrent nets does not guarantee consciousness, destroying the basic postulate.

Nor can the intralaminar nucleus of the thalamus be appealed for unity of consciousness (in whatever sense that is taken), as is clearly seen from those with localised loss of cortex, such as in neglect, but with a perfectly functioning intralaminar nucleus. It is like claiming that all the computation being carried out in my computer as I write this is due to the main generator. Perish the thought!

As for the explanatory gap, the use of no-faster-than light travel based on special relativity is incorrect; there have been numerous attempts to discover "tachyons" which are supposed to travel faster than light. Their existence is not forbidden by Einstein's theory. However, I agree that we should start from the brain to look for the neural mechanisms that create consciousness; I have tried to do that in a forthcoming article (Taylor 1998b). In this I suggest that consciousness arises initially due to local recurrence. The "bubbles" of activity thereby produced by an input, and continuing their existence after the input is switched off, explain the second or so continued posterior neural activity, and persist long enough to fit data such as that of Libet et al. (1964), as explained in Taylor (1996). This is not instantaneous: the dynamics is important. Both vehicle and representation aspects come together here in the crucible of creation of consciousness. The bubbles are suggested as the basis of qualia. The basic properties of the latter can indeed be closely identified with those of the former. This creation of qualia then leads to higher levels of consciousness, with self and thinking now coming into the frame. Frontal lobes then swing into action; consciousness is at least three-tiered (with pre-, passive, and active forms). I suggest that the authors consider how this complexity can arise from neural nettery as the real target. But then they must become neuroscientists!

Quantities of qualia

Michael S. C. Thomas and Anthony P. Atkinson

Psychology Group, King Alfred's College, Winchester SO22 4NR, United Kingdom. **michael.thomas@psy.ox.ac.uk** **atkinsona@wkac.ac.uk** **www.wkac.ac.uk**

Abstract: We address two points in this commentary. First, we question the extent to which O'Brien & Opie have established that the classical approach is unable to support a viable vehicle theory of consciousness. Second, assuming that connectionism does have the resources to support a vehicle theory, we explore how the activity of the units of a PDP network might sum together to form phenomenal experience (PE).

O'Brien & Opie (O&O) claim that there is no room for unconscious explicit representations in a classical vehicle theory. But classicism is saturated with such representations: indeed, it depends on them, as Dennett himself makes clear (1982, p. 218): "So far as cognitive science is concerned, the important phenomena are the explicit unconscious mental representations." We are puzzled as to why O&O think that, in order to advance a vehicle theory, classicists must ground a distinction between conscious and unconscious states on a distinction between explicit representation and representation that is potentially explicit or tacit. It is surely open to the classicist to distinguish two types of explicit representation, and to propose that one of those types generally features in conscious states, and the other in nonconscious states.

Of course, to justify two types of explicit representations without begging the question, we would have to find a criterion such that some explicit representations get to be conscious but others do not. Such a criterion is hard to ground, and for all we know, the cognitive and brain sciences may not yet possess any candidate properties that will stand scrutiny. Perhaps this is why O&O overlook it. But just because it seems difficult to ground the criterion now, does not mean that it will always seem difficult (O&O's argument of sect. 4).

We could take the distinction between conscious and unconscious explicit representations as a given, in the way that O&O take the conscious nature of stable PDP representations as a given. Or we could cast around for a candidate: there is O&O's own notion of *stability;* Farah and colleagues have suggested the notion of *quality* (Farah 1994; 1994b; Farah et al. 1993). Thus stable/high quality explicit representations get to be conscious but unstable/low quality explicit representations do not. What could such terms mean in regard of symbolic representations? Are variables not fully bound? Are representations syntactically poorly formed? PDP seems better equipped to deal with such graded notions. Nevertheless, the classical account is in principle powerful enough to support a vehicle account. It merely requires a distinction between conscious and nonconscious explicit representations based on what those representations are, rather than what they do (sect. 1, para. 7).

We turn now to the second part of our commentary. In O&O's paper, one of the more prominent themes is that of addition. Consciousness is described variously as an "aggregate," an "amalgam," a "sum," "composite," "tapestry," and "multiplicity." Phenomenal experience (PE) emerges from the activity of many simple computational units. We would like to explore a little further exactly how the activity of simple computational units might sum to produce PE.

Consider Equation 1. This says that a single, unified PE is the sum of the activity of units *0* to *i*. Note that the sum has a "magical equals sign" that converts the objective left hand side of the equation to the subjective right hand side.

$$\sum_i a_i = PE \qquad (1)$$

Here are some questions:

1. *Is the sum additive,* so that the more units that are active, the greater the conscious experience? If the sum is additive, would two half-active units produce the same PE as one fully active unit? After all, it is not what units do that is important, just that they are part of a stable pattern.

2. *Does the sum have a threshold?* If so, how many active units are required for consciousness? How much "color" activation must there be for redness to be experienced?

3. *How do the units qualify to be in the sum?* O&O suggest the criterion of stability. This notion needs to be explored further. Activities have to be stable enough, but stable enough for what? For another process to use? Then we have a hybrid vehicle/process account. Stable enough to be conscious? Then we have a tautology.

O&O cite Mangan (1996) as accepting units that have unstable activations. Perhaps we could have an intermediate position, where the sum involves some integral over duration and amount of activation?[1] O&O further cite Lloyd (1993b) as proposing that only *hidden* units should qualify.

Ideally, we would want the question of qualification conditions to be settled empirically. Is there any hope of doing this? At least we might agree to exclude inactive units from the sum. Note that in a process account, this would be unwise since inactive units are informative. For example, the pattern 10010 is defined as much by the zeroes as the ones.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

4. *Is the sum weighted,* so that some units contribute more to the sum than others? (see Equation 2, where X, Y, and Z are the weightings). Perhaps units in more sophisticated circuits could count for more? Perhaps attention could weight the activity of some groups of units over others?

$$X \times \sum_i a_i + Y \times \sum_j a_j + \ldots\ldots + Z \times \sum_k a_k = PE \tag{2}$$

5. *Is there more than one sum?* (see Equations 3–5). Is this view, there is no unitary PE but a diverse set of PEs, each of which is the consequence of a different sum. This is consistent with O&O's position, where distinct modules produce disparate PEs. If there is one sum per module, we may ask "What is the principled distinction such that activity *within* a module is summed, but activity *across* modules is not summed?" In a process account one could simply argue that units in a single module will fall into a single sum because they are working on the same process; units working on different processes will fall into different sums. A vehicle account does not have this option. Finally, note that in the multiple sums account, attention will have to modulate directly the outcome of quite separate sums.

$$\sum_i a_i = PE_1 \tag{3}$$

$$\sum_j a_j = PE_2 \tag{4}$$

$$\sum_k a_k = PE_N \tag{5}$$

NOTE

1. Thanks to Bob French for this idea.

Vehicles, processes, and neo-classical revival

Robert Van Gulick

Department of Philosophy, Syracuse University, Syracuse, NY 13244-1170.
rnvangul@mailbox.syr.edu

Abstract: O'Brien & Opie unfairly restrict the classicist's range of options for explaining phenomenal consciousness. Alternative approaches that rely upon differences among representation types offer better prospects of success. The authors rely upon two distinctions: one between symbol processing and connectionist models, the other between process and vehicle models. In this context, neither distinction may be as clear as they assume.

According to O'Brien & Opie (O&O), the classical computational theory of mind lacks the resources to develop a vehicle theory of phenomenal consciousness. Their argument, however, unfairly restricts classicism's explanatory options. They invoke Dennett's (1981) four-fold distinction among modes of representation (explicit, implicit, potentially explicit, and tacit) and require the classicist to locate conscious representation uniquely within that scheme. They plausibly conclude that no such necessary and sufficient condition can be given. One cannot, for example, count all explicit representations as conscious. But that shows at most that classicism cannot construct a vehicle account of consciousness using only Dennett's distinctions. It does not show that it lacks the further requisite tools.

A classicist might appeal to the features of specific types of representations to help explain the nature and basis of conscious thought. Phenomenal mentality has many distinctive features, such as the globally integrative nature of its content, its perspectival presentation from the focus of a unitary self, and the apparent sensuous manifolds associated with many of its modalities. The classicist has a lot more hope of explaining such features in terms of the representation types that subserve them than by appeal to Dennett's four-way scheme. Thus insofar as O&O ignore classicism's most promising option, their negative conclusion about its prospects seems at best premature and in need of further argument.

O&O's own connectionist proposal relies upon the same four-way framework of distinctions they impress on the classicist. It claims that generating an explicit activation pattern representation in the brain is both necessary and sufficient for phenomenal consciousness. Whatever advantages this proposal may have over the classical options they reject, it remains implausible, and largely for the same reason: it asks the explicit/implicit/tacit distinction to do more than it can. Even if activation patterns play a role in human consciousness, it is unlikely that just any activation pattern will suffice; it is too easy to get a neural network into a stable pattern of activation. Such patterns can and probably do occur throughout the brain in functional roles that rule them out as candidates for consciousness, such as when they occur early in the perceptual process. Even on O&O's generous views about the multitudes of representations simultaneously conscious in one's mind, many stable activation patterns still would not qualify. They simply do not occur at the right stage of processing or have the right sort of content. If the content of a stable activation pattern concerns a section of the retinal light array, then no view of conscious states – no matter how generous – can count it. Our conscious phenomenal life just does not include such contents.

Nonetheless, the prospects for a connectionist theory of consciousness look bright. Activation patterns of the specific types discussed by O&O seem well suited to serve as the neural substrate for important features of phenomenal experience. Following Churchland (1995) and others, O&O argue that relations of phenomenal color similarity can be modeled by activity across a hyperdimensional activity space. However, the explanatory value in such cases derives not from the fact that the vehicles of consciousness are activity patterns per se but from their particular natures, and the ways in which those dynamic or structural features can be put into correspondence with parallel aspects of experience. In this respect, the general explanatory strategies for the connectionist are similar to those that are most promising for the classicist.

This explanatory convergence suggests that some of the distinctions on which O&O rely in staking out their position may not be as clearcut as they suppose. First, one might aim for a more ecumenical approach that treats the connectionist story as implementing a classical symbol processing model. The stable activation patterns could be treated as the symbolic structures, with the processes governing their interaction embodied in the associated intra- and inter-network linkages. One can propose a symbol processing model without supposing that the relevant symbols are simply digital strings or even less likely characters on a Turing machine's tape; indeed, there is no reason why they could not be global integrated structures built from stable or self-reinforcing patterns of activation. Consider, for example, Kinsbourne's (1988) integrated field theory of consciousness or Flohr's (1991) theory of phenomenal consciousness in terms of transient large scale neural assemblies. Both models have a lot in common with O&O's proposals, yet each is also compatible with a symbol processing outlook as long as one does not take an over-restrictive view of what can count as symbols or as processing them.

The second distinction that may bend a bit under pressure is the one between vehicle and process models of consciousness. O&O put a lot of weight on this and stress the alleged general absence of vehicle models as opposed to process ones. But once attention shifts away from the implicit/explicit/tacit distinction and focuses on the specific types of representations, the line between the two sorts of models begins to blur. Type-based models focus on the differences among specific sorts of representations (vehicles) and the diverse interactions into which they can enter (processes). Indeed, the most important differences among representations will prob-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

ably concern those features that affect their ability to enter into differing roles and interactions. Put in a slogan, vehicle differences matter when they make a difference to processes. If that is so, any interesting vehicle theory of consciousness will of necessity also be a process theory of consciousness. There is no need to choose between the two, and indeed perhaps no real way to do so.

Brute association is not identity

Bram van Heuveln and Eric Dietrich

Program in Philosophy, Computers, and Cognitive Science, Binghamton University, Binghamton, NY 13902-6000.
bram@turing.paccs.binghamton.edu
dietrich@turing.paccs.binghamton.edu www.paccs.binghamton.edu

Abstract: O'Brien & Opie run into conceptual problems trying to equate stable patterns of neural activation with phenomenal experiences. They also seem to make a logical mistake in thinking that the brute association between stable neural patterns and phenomenal experiences implies that they are identical. In general, the authors do not provide us with a story as to why stable neural patterns constitute phenomenal experience.

We have two problems with O'Brien & Opie's (O&O's) target article. Our main concern is that the proposed theory does not appear to be one about phenomenal experience at all. Perhaps it can be seen as a theory of attention. However, regarded as such, the paper reveals a second, fundamentally logical, mistake.

1. O&O's main hypothesis is that "phenomenal experience consists in the explicit representation of information in neurally realized PDP networks." Later we read that this explicit representation is to be understood as a stable pattern of neural activation. Our problem is that we do not see how consciousness can consist in stable patterns of neural activation. We can all observe stable patterns of neural activation in the brain of someone else. However, no one is able to observe the phenomenal experiences of this particular person. Hence a phenomenal experience cannot be the same as some stable pattern of neural activation.

Some philosophers have argued that stable neural patterns and phenomenal experiences could very well be two different perspectives, two different "ways of knowing" the same entity (see Churchland 1995). O&O, however, do not make such a claim. Their claim is that neural patterns are identical to phenomenal experiences, and that position is ruled out by our argument above. As O&O themselves point out, any materialist theory runs into the "hard" problem of consciousness. Indeed, although the authors devote a whole section (The Explanatory Gap) to the "hard" problem, we find their reasons for believing that their own materialist theory fares any better are unconvincing:

(a) O&O state that their theory is not worse off than any other theory. This is compatible with our view that theirs, like any other materialist theory, has nothing to offer when it comes to the "hard" problem.

(b) Since we can conceive of creatures that have stable neural patterns of activation but no phenomenal experiences, O&O's theory seems implausible. Their defense here is that scientific investigation can change our conceptions over time, and hence also our ability to conceive of such creatures. Our problem with this reply is that explanations lie in the present, and not in the future. Right now, stable patterns of neural activation are clearly conceived as entities distinct from phenomenal experiences, so right now O&O's theory is unsatisfactory. One cannot appeal to our future conceptions to make a currently unsatisfactory theory in any way attractive.

(c) Finally, O&O state that phenomenal experiences are complex entities having many structural and temporal properties, and that their own connectionist vehicle theory has the potential to model all these similarities and differences between phenomenal experiences. O&O believe that the more such similarities we find between stable neural patterns of activation and phenomenal experiences, the closer we come to closing the explanatory gap. However, although it may be true that neural patterns can indeed *mirror* all the complex details of phenomenal experience, we do not see how this would make the two *identical.* This holds even if one day someone developed a completed connectionist vehicle theory that had a total mapping between all possible stable neural patterns and phenomenal experiences. In such a case, there would be good reasons to believe that any creature with certain specific stable patterns of activation would have certain specific phenomenal experiences, but it would still be mysterious why this should be the case. A theory that cannot explain why phenomenal experiences *are* stable patterns of activation is not the theory of phenomenal experience that philosophers and cognitive scientists are looking for.

2. Perhaps O&O's connectionist vehicle can be seen as a theory of what psychologists study when they study states of attention. States of attention are the objectively observable properties of consciousness. Indeed, as opposed to phenomenal experience, attention is open to scientific investigation, and psychologists have come up with working definitions to study, measure, and quantify it.

How plausible is O&O's theory as a theory of attention? Our personal intuition is that stable neural patterns do not constitute states of attention in and of themselves. We think it much more appropriate to analyze a cognitive agent's currently attending to something as some complex process involving a lot more of the agent and its environment, rather than some isolated stable neural pattern solely within that agent. So, we would opt for a version of a process theory of attention rather than a vehicle theory. Do O&O's arguments provide any reasons to convert to a vehicle theory of attention? We think not, and here is why.

If we view O&O's theory as a theory of attention, we can recast their arguments as making a case for the view that attention equals stable patterns of neural activation because attention shares certain essential structural properties with such stable patterns. But having shared essential properties is not enough to warrant the conclusion that attention is identical to stable patterns because there can be reasons other than their being identical that explain why attention and stable patterns have so much in common. For example, the fact that one's car makes a left turn whenever one turns the wheel to the left does not make the two actions identical, even though both actions share the property of turning to the left, and despite the fact that there is a law-like relationship between turning the wheel to the left and the car going to the left.

Even inductively, the conclusion that attention is stable patterns of neural activation is no more plausible than the view that attention is a larger process, perhaps involving some sort of executive awareness focussing system, in which those stable patterns only play a part. The fact that stable patterns of neural activation have so much in common with states of attention equally supports both views; O&O do not provide us with any further evidence in support of their vehicle theory and against the process theory.

In sum, the conceptual problems that arise from trying to equate stable neural patterns with phenomenal experiences (or, for that matter, from trying to equate anything physical with phenomenal experiences) cry out for a *story* as to why stable patterns of neural activation would be identical to phenomenal states, a story that goes beyond a mere law-like relationship. Ideally, O&O should have given us analyses of the concepts involved, followed by such a story. But as we said, we don't think such a story is in the offing, certainly not in the foreseeable future. O'Brien & Opie, like everyone else, are left with a brute association between physical states (in their case stable neural patterns) and phenomenal experiences. Thus, we are left in the dark as to how any physical state could be some phenomenal state.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Neural activation, information, and phenomenal consciousness

Max Velmans

Department of Psychology, Goldsmiths, University of London, London SE14 6NW, England. **m.velmans@gold.ac.uk**
www.gold.ac.uk/academic/ps/velmans.html

Abstract: O'Brien & Opie defend a "vehicle" rather than a "process" theory of consciousness largely on the grounds that only conscious information is "explicit." I argue that preconscious and unconscious representations can be *functionally* explicit (semantically well-formed and causally active). I also suggest that their analysis of how neural activation space mirrors the information structure of phenomenal experience fits more naturally into a dual-aspect theory of information than into their reductive physicalism.

It is self-evident that something in the brain must differentiate conscious from preconscious and unconscious states. In their thoughtful article, O'Brien & Opie (O&O) suggest that conscious states are characterised by stable (versus unstable) patterns of activation in neural networks – a physical "vehicle theory" of consciousness in which each phenomenal experience is identical to a stable pattern of neural activation. Their argument in favour of a vehicle theory rather than a classical "process" theory largely centres on the claim that only conscious information is *explicit* (is formed into physically distinct, semantically interpretable objects) – and a stable activation pattern is appropriately explicit. Classical processing theories involving symbol manipulation assume that much *nonconscious* information is also explicit (in which case something has to be *done* to the information to make it conscious). Neural nets, they suggest, combine explicitness and consciousness in a more natural way. Given its potential for advancing our understanding of the physical substrates of phenomenology, their case merits serious consideration.

Much depends, of course, on whether *only* conscious information is explicit. Given the massive evidence that at least some preconscious and unconscious information is "explicit" (in the sense of being sufficiently well-formed to be semantically interpretable), O&O's claim requires *all* the evidence to the contrary (not just some of it) to be methodologically flawed – and it is notable that in making their case they rely on a strictly one-sided reading of the literature (for example, they cite reviews by Holender 1986 and Shanks & St. John 1994, but ignore extensive, contrary reviews by Dixon 1971; 1981; Kihlstrom 1996; Reber 1997; and Velmans 1991). Even *one* good example of preconscious or unconscious semantic processing would be troublesome for their theory and there are many examples which, to my knowledge, *have never been challenged.* Groeger (1984) for example, found evidence of preconscious semantic analysis in a nonattended ear, under conditions that cannot be explained by focal-attentive switching (with accompanying consciousness). That is, he found that the effects of disambiguating words in the nonattended ear on a sentence completion task in the attended ear were different if the nonattended words were at threshold (consciously detectable) versus below threshold.

For example, in one experiment subjects were asked to complete the sentence "She looked ____ in her new coat" with one of two completion words, "smug" or "cosy." Simultaneous with the attended sentence the word "snug" was presented to the nonselected ear (a) at threshold, or (b) below it. With "snug" presented at threshold, subjects tended to choose "smug," which could be explained by subjects becoming momentarily aware of the physical form of the cue. With "snug" presented *below threshold,* subjects tended to choose "cosy," indicating semantic analysis of the cue without accompanying awareness. That is, below-threshold, nonattended, semantic information can be causally active – and according to O&O (sect. 3.2, para. 3) that makes it explicit.

Other experiments show that when spoken words *are* attended to, their multiple meaning are simultaneously, preconsciously activated (in the first 250 milliseconds). Depending on context, one meaning is selected and the subsequent entry of the word into consciousness is accompanied by inhibition (or deactivation) of inappropriate meanings (Pynte et al. 1984; Swinney 1979; 1982). Such briefly activated, preconscious, semantic codes give every appearance of being sufficiently well-formed to influence subsequent processing, as classical theory suggests. Long-term memory provides an additional store of encoded meaning, comprising our knowledge of the world. Such knowledge is largely *unconscious* and stable, although it is causally active in determining our expectations and interactions with the world. O&O suggest that in PDP systems this can be handled by the connection weights and patterns of connectivity (sect. 4.1, para. 12). But, in a sense, the "vehicle" which caries this information is irrelevant to whether it is unconscious, causally active and *functionally* "explicit." If a waiter gives one the bill before the menu, one knows *immediately* that something is wrong – one does not have to consciously rehearse a script of what is supposed to happen in restaurants! So, even if O&O are right, such unconscious "connection weight representations" must be sufficiently "explicit" (semantically well-formed) to act as they do.

O&O's physicalist reductionism also needs to be treated with caution. They take it for granted that if "vehicle" theory is correct, then, "the complex physical object constituted by the stable pattern of spiking frequencies is the phenomenal experience" (sect. 5.1, para. 8). Nowhere in their target article, however, do they bother to defend this *ontological identity* claim. A neural activation "vehicle" is a *carrier* of information. If O&O are right, such activation patterns *correlate* with phenomenal experience – and, in section 5.4, they give an interesting analysis of how similarities and differences in the "dimensionality" and "shape" of neural "activation spaces" might *mirror* patterns of similarity and difference in phenomenal experience. The necessary and sufficient conditions for the creation of such "activation spaces" could also then be thought of as the *causes* of phenomenal experience. But *correlation* and *causation* are very different from *ontological identity* (cf Velmans 1998).

I do not have space to elaborate on these distinctions here. But it should be clear that while "information structure" can express the patterns of similarity and difference in phenomenal experience, it does not capture its "subjectivity" and "qualia." One might, for example, know everything there is to know about the "shape" and "dimensionality" of a given neural activation space and still know nothing about what it is like to have the corresponding experience. This is obscured in the normal, human case by the fact that *third-person* access to brain states is complemented by *first-person* access to our own experience. By means of this dual access, we can discover whether certain "activation spaces" correspond to "auditory experiences," others to "visual experiences," and so on. If silicon had the appropriate "qualia producing" powers, it might then be possible to construct neural nets with the same "activation spaces" and corresponding experiences. But suppose we arrange a net to operate in a nonhuman configuration, with an "activation space shape" which is quite unlike that of the five main, human, sensory modalities. What would it experience? We cannot know! And here's the point: if we can know the "shape" of the space very precisely and still do not know what it is like to have the experience, then having a particular activation space cannot be *all there is to having an experience*!

Such points (which echo Nagel 1974) are very difficult to accommodate within a reductive "physicalism" or "functionalism" which tries to translate the phenomenology of first-person experience entirely into how things appear from a third-person point of view, although they present no impediment to nonreductive positions. O&O's analysis of how the information structure of neural activation space mirrors that of phenomenal space fits naturally, for example, into a dual-aspect theory of information (of the kind that I have proposed in this journal in Velmans 1991; 1993; 1996). This accepts that information encoding in the brain, PDP systems, and so on can only be properly known from a third-person perspective, while phenomenal experience can only be properly

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

known from a first-person perspective. The patterns of similarity and difference ("the information structure") within a given phenomenal experience and its neural correlates is identical, but this information appears in very different neural and phenomenal formats for the reason that first- and third-person ways of accessing that information (the "observational arrangements") are very different. A shared information structure allows one to *relate* first-person phenomenology to third-person neural accounts very precisely, but it does not "reduce" the phenomenology to "activation space" (or to any other physical correlate). On this view, first- and third-person observations of consciousness and brain are complementary and mutually irreducible. A complete account of mind requires both.

What about consciousness during learning?

Annie Vinter and Pierre Perruchet
L.E.A.D., CNRS 5022, Faculty of Sciences, 21000 Dijon, France.
vinter@ubourgogne.fr

Abstract: Though we fully agree that unconscious processing produces explicit representations that form the conscious phenomenal experience of the subject, identifying phenomenal experience with stable patterns of activation in a PDP network seriously limits O'Brien & Opie's thesis. They fail to recognize the constructive role of consciousness during the learning episode itself, reducing consciousness to a resulting outcome of the learning episode. We illustrate how consciousness can guide and shape the formation of increasingly structured representations of the world by presenting a brief outline of a model for speech segmentation.

One of O'Brien & Opie's (O&O's) main theses is that "explicit representations . . . are the products of unconscious processes" and that identifying "phenomenal experience with the vehicles of explicitly representation in the brain" construes consciousness as a "fundamental feature of cognition." This thesis is clearly not in line with the dominant Zeitgeist in cognitive psychology, which largely discards consciousness and phenomenal experience from its main explanatory concepts. However, it finds some echo in our own account of implicit learning (see, for instance, Perruchet & Vinter 1998). In our account, unconscious *processes* and conscious *representations* are conceived as the front and the reverse of a sheet of paper. As the analogy illustrates, they are both intrinsically associated and radically different: radically different, in the same way that any overlap between the front and the reverse is obviously impossible; and intrinsically associated, any dissociation between the front and the reverse being likewise impossible. In implicit learning, intrinsically unconscious processes serve the function of generating conscious representations, hence shaping the phenomenal experience of the subject. Noticing the similarity of O&O's position and ours seems important, especially because both views stem from very different backgrounds. O&O's is a philosophical approach to connectionism, whereas we rely on experimental and developmental psychology, without any commitment to a connectionist perspective.

However, we no longer follow O&O when they identify each phenomenally experienced representation with the generation of a stable pattern of activation in a PDP network. The authors do not define clearly what they mean by a "stable" pattern of activation in neurally realized PDP networks. We can infer that they mean the final relaxation state presented by a net after training, when activity is fully stabilized, linking inputs and outputs in a coherent way. If we take this interpretation for granted, this entails a radical dissociation between phenomenological experience and learning. Indeed, learning is linked with the period during which the weights of the net adjust themselves while each new input is processed, and phenomenological experience emerges when the weights no longer change.

This dissociation raises an obvious problem. O&O's position leads to the paradoxical claim that there is no phenomenal experience during learning. Consider, for example, a connectionist modelling approach to speech segmentation such as Elman's (1990) SRN model, the objective of which is to reproduce the human ability to correctly segment a continuous speech stream into words. Activity in the net will be fully stabilized when the net has learned to segment the utterances correctly. But what about the phenomenal experience of the input at the beginning of the presentation of the linguistic corpus, before the relaxed states are achieved? It is obvious that a human subject phenomenally experiences the perceived input, even while appropriately structured representations are not yet available.

Moreover, because consciousness is not introduced during the learning episode itself, O&O fail to recognize the possible role played by the initial conscious representations in the formation of the ultimate representations. Consciousness appears as a terminal or final state of what has been learned by the net, without the "vehicles of explicit representations" having an active role in the formation of the subsequent explicit representations. Consciousness is a fundamental feature of cognition in the sense of a final or resulting feature, not in the sense of a constructive feature. In our view, the conscious explicit representations forming the momentary phenomenal experience of the subject play an active role in the process of formation of subsequent, better structured explicit representations.

The way initial, poorly structured conscious representations may contribute to learning can be illustrated by a brief outline of the principles of PARSER (see Perruchet & Vinter, in press, for a detailed presentation), a nonconnectionist model for speech segmentation. We started from the consideration that, faced with a continuous speech stream in an unknown language, humans naturally segment this information into small and disjuntive "chunks," each chunk embedding a few primitives. In PARSER, this initial parsing is simulated by a random generator. The chunk, or percept, forms the content of the subjects' momentary phenomenal experience. It also enters as a unit in memory, and is ascribed a weight. This weight is increased if the chunk is perceived again, and decreased from a certain quantity to simulate forgetting, and possibly interference, each time another percept is processed. Crucially, as long as the weight of a memory unit is above a certain threshold, this unit has the property of guiding perception. Thus, the new conscious units progressively substitute for the primitives of the system. As a consequence, when a chunk already stored in memory is present in the input, it will be perceived as a unitary percept, instead of being cut off in several parts. This makes the model very efficient for extracting the regularity from the input. As a matter of fact, after some training, most of the items present in memory are the words of the language, because the probability of drawing the same chunk (or encountering the same percept) repeatedly is higher if this percept is a word, or a part of a word, than if it straddles word boundaries. Thus in PARSER, the words emerge through some kind of natural selection process, the nonwords being progressively forgotten because too rarely repeated.

The point is that, in our model, the conscious representations are not only the final products of learning, as in O&O's theory: they are present as the very beginning of training and serve throughout the learning process, thanks to their ability to constrain the coding of the incoming information. In Piagetian terms, ascribing a role for phenomenal consciousness in the formation of structured representations allows our model to re-integrate "assimilation," along with "accommodation," in adaptive processes.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Constructing consciousness

Gezinus Wolters[a] and R. Hans Phaf[b]

[a]*Department of Psychology, Leiden University, 2300 RB Leiden, The Netherlands;* [b]*Department of Psychology, University of Amsterdam, 1018 WB Amsterdam, The Netherlands.* **wolters@rulfsw.leidenuniv.nl**
pn_phaf@macmail.psy.uva.nl

Abstract: O'Brien & Opie make unnecessary distinctions between vehicle and process theories and neglect empirically based distinctions between conscious and unconscious processing. We argue that phenomenal experience emerges, not just as a byproduct of input-driven parallel distributed processing, but as a result of constructive processing in recurrent neural networks. Stable network states may be necessary, but are not sufficient, for consciousness.

For a distinction such as that between vehicle and process theories to be meaningful, it should be shown that the two poles, if not mutually exclusive, do not almost always co-occur. O'Brien & Opie (O&O) acknowledge that they can be combined (note 5) but dismiss this as unparsimonious. It is our contention that many accounts of consciousness, including O&O's, are "hybrid" in this respect. O&O identify phenomenal experience with stable states arising in neural networks through relaxation on a coherent set of activations (Kihlstrom 1987). In connectionist relaxation, a network selects from among all possible activation states the one that satisfies most constraints (i.e., connections) activated by the input. Potentially explicit and tacit information can thus be causally active in determining these stable patterns. This, however, also adheres to O&O's definition of a process theory: "conscious experience is the result of a superordinate computational process or system that privileges certain mental representations over others." O&O's characterization of Baars's (1996) global workspace theory may be paraphrased here: the nature of the vehicles is secondary; what counts, so far as consciousness is concerned, is relaxation into a stable state.

A more important distinction than between vehicle and process theories is between conscious and unconscious processing. If consciousness is merely a by-product of nonconscious processing (the identity position, Mandler 1985), the implementation of nonconscious processing would suffice to model consciousness. Only when qualitative differences occur between the two is it useful to postulate separate vehicles or processes for consciousness. O&O's description of dissociation research does not do full justice to the findings in this field. They ignore a vast amount of evidence for qualitative differences from studies on implicit perception (Greenwald 1992; Merikle 1992), implicit memory (e.g., Roediger & McDermott 1993) and nonconscious affective processes (LeDoux 1996; Murphy & Zajonc 1993). This research, for example, contradicts the idea that subliminal presentation corresponds only to a diluted form of conscious processing. The neglect is reflected in O&O's connectionist theory. The networks they describe only react to input, they have no capability of actively manipulating information. Involuntary (stimulation-caused) and voluntary (expectation-caused) forms of selective attention clearly have to be distinguished, but only the former figure in O&O's theory. Moreover, other aspects of phenomenal experience, such as the symbolic and sequential nature of conscious contents and the role of productivity, remain unaddressed.

Although we strongly endorse a connectionist approach to consciousness (Phaf et al. 1994; Phaf & Wolters 1997), we believe that O&O's story is incomplete and ignores the functions of consciousness. In our view, network relaxation provides the ingredients for subsequent constructive processing (Mandler 1985), which is ultimately responsible for phenomenal experience. Constructive processes are needed for the appraisal of situations, for creating models of the world and expectations about the outcome of actions, and for performing the recurrent operations in planning, thinking, and problem solving. To allow such functions, constructive processes have to meet the following requirements. First, the system should be able to operate on all representations derived from relaxation processes. Second, contents of constructions should be able to guide subsequent relaxations (i.e., attention can be redirected). Third, these contents should have a symbolic format (relaxation transforms subsymbolic network-activations into – for example, verbal – output). Fourth, the contents should be able to contact almost anything stored in the network (allowing temporary couplings of representations not directly associated in long-term memory). From these requirements it can be deduced that construction processes occur in a specific version of working memory.

Perhaps O&O would call our view a process theory, because it seems to emphasize the processes involved in constructing phenomenal experience, but we also put some restrictions on the type of experience (i.e., vehicles) to be constructed. The version of working memory we envisage is not some unitary executive system (Churchland 1995), or an executive with slave systems (Baddeley 1986), or a global workspace (Baars 1996). Instead, we believe that in evolution several working memories arose through the internalization of the external loop of "object perception – response – perception of result." By internalizing such loops, maintaining and manipulating symbolic representations is no longer limited by the physical presence of objects or situations. In the human system, internal modelling may be possible in the auditory-articulatory (e.g., verbal descriptions), the visuo-spatial (e.g., imagery), and the somato-sensory (e.g., experiencing bodily states) domains. A network model of working memory capable of holding and manipulating symbolic representations through sequential feedback (cf subvocal rehearsal) is thus implemented by the combination of a multimodular network with several recurrent loops.

We agree that connectionism provides a better opportunity to model consciousness than classical computational theory (Greenwald 1992; Kihlstrom 1987). We also acknowledge that O&O do a laudable job in their discussion of knowledge representation in classicism and connectionism. The target article provides insights into the role of nonexplicit knowledge in information processing, and the idea that stable states are to be seen as a complex amalgam of nonconscious elements derived from simultaneous constraint satisfaction. Simply equating stable states with conscious experience, however, underestimates empirical evidence for dissociations and ignores the evolutionary adaptive functions of consciousness. To close the "explanatory gap," we need more attempts at working models for empirical data (even if not completely successful ones) and fewer verbal and introspective arguments. O&O's theory provides starting points for connectionist models of consciousness, but it should be supplemented both by additional processes and restrictions to the kinds of possible vehicles for phenomenal experience.

Priming in neglect is problematic for linking consciousness to stability

Marco Zorzi[a] and Carlo Umiltà[b]

[a]*Dipartimento di Psicologia, Università di Trieste, 34123 Trieste, Italy;* [b]*Dipartimento di Psicologia Generale, Università di Padova, 35131 Padua, Italy.* **zorzi@univ.trieste.it umilta@psico.unipd.it**
www.psychol.ucl.ac.uk/marco.zorzi/marco.html

Abstract: O'Brien & Opie argue that (1) only explicit representations give rise to conscious experience, and (2) explicit representations depend on stable patterns of activation. In neglect patients, the stimuli presented to the neglected hemifield are not consciously experienced but exert causal effects on the processing of other stimuli presented to the intact hemifield. We argue that O'Brien & Opie cannot account for a nonconscious representation that is stable, as attested by the fact that it affects behavior, but is neither potentially explicit nor tacit.

O'Brien & Opie (O&O) propose stability as the central feature of their connectionist theory of phenomenal experience. In their

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

view, information is explicitly coded only by stable patterns of activation. In contrast, the activity prior to stabilization (such as, for example, the process of settling into an attractor) does not produce explicit coding and thus does not produce phenomenal experience either.

A fundamental postulate of the theory is that explicit representation of information in the brain and conscious experience cannot dissociate. Therefore, O&O argue that dissociation studies (e.g., blindsight, implicit learning, etc.) do not show such a dissociation conclusively. In the case of neuropsychological dissociations, their reasoning seems to apply to blindsight but not, for example, to neglect. An important source of constraints for a theory of consciousness is what is usually referred to as "implicit processing" in neglect patients. The results of neglect studies were not even mentioned in the target article. The behavior of neglect patients has important theoretical implications which do not easily fit into O&O's framework.

Patients with unilateral neglect after (right) parietal lesions ignore the affected (left) side of space (e.g., papers in Robertson & Marshall 1993). They behave as if the left half of the world had ceased to exist at a conscious level. There is considerable evidence, however, that neglect patients show normal processing of neglected stimuli in the affected left side. For example, Berti and Rizzolatti (1992) presented pictures of objects to the neglected hemifield of patients with severe neglect. Although the patients were completely unaware of these stimuli, the pictures primed (i.e., facilitated) responses to semantically related target objects presented to the intact hemifield. This effect was also present when the prime and the target were physically different but belonged to the same category.

A similar phenomenon was documented by Làdavas et al. (1993). Their patient could not read aloud words presented to the left hemifield, nor could he judge (or "guess") their semantic content or lexical status. He could not even detect the presence of letter strings in that hemifield. However, response to a written word in the intact hemifield was faster when the word was preceded by a brief presentation of an associated word in the neglected hemifield.

How does the pattern of behavior shown by neglect patients fit into O&O's scheme? Clearly, the patients have no phenomenal experience of the stimuli presented in the neglected hemifield. Yet, these stimuli are processed, in a seemingly normal fashion, to the point that they exert *causal* effects on the processing of other stimuli. However, information is not explicitly represented (otherwise it would be consciously experienced), so it must be coded in a non-explicit fashion. According to O&O, two kinds of non-explicit representations are possible within their vehicle theory of consciousness: (1) potentially explicit representations, and (2) tacit representations. Tacit representations can readily be excluded, because in a PDP framework they concern processes such as activation and output functions of the individual units (neurons). Potentially explicit representations, on the other hand, are discussed by O&O in terms of connection weight representation: that is, the information encoded in the weight matrix is potentially explicit because it can be rendered explicit by an appropriate input vector.

Can potentially explicit representations explain the behavior of neglect patients? The stimuli presented to the neglected hemifield exert causal effects on the processing of other stimuli presented to the intact hemifield (the more typical effect taking the form of priming). Thus, given that the patient studies showed *semantic* priming, one has to conclude that the neglected stimulus is processed in the brain to the point at which a semantic representation of the object is produced. Most theories of semantic priming (e.g., Collins & Loftus 1975; see review in Neely 1991), including connectionist models (e.g., Masson 1995; Plaut 1995), assume that priming effects are the results of *activation* produced by prior processing of the prime when the target stimulus is presented. In connectionist models, concepts are represented as distributed patterns of activity over a large set of processing units that encode microfeatures. Related concepts are represented by similar overlapping patterns (e.g., McRae et al. 1997). Semantic priming is produced by the overlap between prime and target. Because processing of the target starts from the activation pattern produced by the prime, it will be faster when the prime is related (i.e., partially overlapping).

All this renders it very unlikely that priming is due to potentially explicit representations based on the connection weights. It would therefore seem that neither of the two non-explicit forms of representation in O&O's taxonomy can be the basis of the behavior shown by neglect patients. Farah (1994) proposed a quality of representation account of "implicit processing" in neglect and other neuropsychological syndromes, in which conscious experience requires a relatively higher quality of perceptual representation than nonconscious perceptual performance does. The representations from the neglected hemifield would be below this quality threshold. In connectionist terms, that might correspond to a representation that fails to become stable.

Finally, we would like to point out an apparent contradiction in O&O's reasoning. Because the dissociability of conscious experience and explicit representation, if proved true, would seriously undermine their theory, they devote much space to arguing that such a dissociation has yet to be adequately demonstrated. Then they invoke those some dissociation studies for maintaining that there is neuropsychological evidence pointing to the distributed neural basis of consciousness. For example, they cite the papers in Milner and Rugg (1992), which deal with "implicit processing" in blindsight, neglect, prosopagnosia, and other neuropsychological syndromes. We share O&O's view that phenomenal experience is not unitary and does not depend on a single neuroanatomical structure (see Umiltà & Zorzi 1995). However, we cannot see how one can reject the neuropsychological dissociations when discussing the link between stability and consciousness, and then use them to support the distributed nature of the neural substrate of consciousness.

ACKNOWLEDGMENT

Preparation of this paper was supported by grants from CNR and MURST.

Authors' Response

Putting content into a vehicle theory of consciousness

Gerard O'Brien and Jonathan Opie

Department of Philosophy, The University of Adelaide, South Australia 5005, Australia. **gobrien@arts.adelaide.edu.au**
chomsky.arts.adelaide.edu.au/philosophy/gobrien.htm
jopie@arts.adelaide.edu.au
chomsky.arts.adelaide.edu.au/philosophy/jopie.htm

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

Our target article sets out to defend a way of thinking about consciousness that, although not completely novel, is cer-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Table R1. *Outline of Response*

Part I: Conceptual foundations	
Section 1. Vehicles	**(Church, Cleeremans & Jiménez, Dennett & Westbury, Kurthen, McDermott, O'Rourke, Thomas & Atkinson, Van Gulick, Wolters & Phaf)**
Section 2. Representation	**(Church, Clapin, Dennett & Westbury, Ellis, Lloyd, Mac Aogáin, O'Rourke, Perner & Dienes, Reeke, Schröder, Wolters & Phaf)**
Section 3. Stability	**(Cleeremans & Jiménez, Dennett & Westbury, Gilman, Lloyd, Mangan, McDermott, Perner & Dienes, Pólya & Tarnay, Reeke, Schröder, Taylor)**
Section 4. Identity	**(Ellis, Kurthen, Newton, van Heuveln & Dietrich, Velmans)**
Part II: Empirical plausibility	
Section 5. General empirical concerns	**(Cleeremans & Jiménez, Gilman, Mangan, Mortensen, O'Rourke, Perner & Dienes, Schröder, Van Gulick, Velmans)**
Section 6. The dissociation studies revisited	**(Dulany, Kentridge, Perner & Dienes, Velmans, Vinter & Perruchet, Zorzi & Umiltà)**
Section 7. Subject unity, agency, and introspection	**(Carlson, Coltheart, Dulany, Mac Aogáin, McDermott, Perner & Dienes, Schwitzgebel)**

tainly unfashionable in contemporary cognitive science. Most theories in this discipline seek to explain conscious experience in terms of special computational processes that privilege certain of the brain's representational vehicles over others. In contrast, we conjecture that phenomenal experience is to be explained in terms of the intrinsic nature of the explicit representational vehicles the brain deploys – in terms of what these vehicles *are* rather than what they *do*. Given that vehicle theories of consciousness are rare in cognitive science, we expected our target article to receive a good deal of criticism, and our expectations have certainly been met. What is gratifying, however, is the constructive spirit in which this criticism is proffered. If, by the end of this reply, our connectionist vehicle theory of consciousness is any more intelligible (and, dare we hope, any more plausible), it is the commentators we have to thank.

The commentaries raise difficulties and objections across many fronts, ranging from the neuroscientific to the philosophical. It has been a daunting task to structure a reply that responds to all of these worries in a perspicuous fashion. In the end, we have settled on seven interrelated themes, embracing both the conceptual foundations of our vehicle theory (Part I), and its general empirical plausibility (Part II; see Table R1).

Part I: Conceptual foundations

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

R1. Vehicles

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

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

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

only a vehicle theory can satisfy one of the deepest intuitions we have about our conscious experience – that it makes a difference.

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

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

But can both of these interpretations of informational access be sustained by process theorists? We think not. It makes little sense to talk of a particular representational vehicle enjoying rich and widespread information processing *relations* in a cognitive system unless it is actually having rich and widespread information processing *effects.* Dennett, we believe, has seen this, and so avoids reading rich informational access in terms of the capacities of a select subset of representational vehicles. Instead, he concentrates on what these vehicles actually *do* in the brain – the impact they have on the brain's ongoing operations. As a result, phenomenal experience, according to Dennett, is like fame, a matter of having widespread effects. With regard to pain, for example, he argues that our phenomenal experience is not identifiable with some internal state that is poised to cause typical pain reactions in the system; rather, "it is the reactions that *compose* the 'introspectable property' and it is *through reacting* that one 'identifies' or 'recognizes' the property" (1993, p. 927). Consequently, in spite of the barrage of criticism that has been leveled at Dennett's account of phenomenal consciousness over the years, his position is actually more consistent with the general entailments of process theories.

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

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

Thomas & Atkinson and **Van Gulick** object to this last claim. They argue that we have failed to consider the possibility that among the classical vehicles of explicit representation there are distinct *types* (i.e., distinct subspecies of symbol structures), one of which might plausibly be aligned with phenomenal experience. What lends weight to this objection is the observation that both computer science and cognitive psychology appear to distinguish between different kinds of explicit representations. We find computer scientists referring to *simple* and *composite* data types, for example, and cognitive scientists referring to *frames, scripts, propositional encodings,* and *productions,* to name but a few. The suggestion is that the classicist might be able to exploit these distinctions to provide a *vehicle* criterion for consciousness. We do not think this will work. The distinctions among structured data types in computer science, and among the various kinds of representations in cognitive psychology, are based on the *differential computational roles* of

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

these vehicles, rather than on their intrinsic physical properties (see O'Brien & Opie, forthcoming, for a detailed defense of this claim). To associate conscious experience with a particular type of explicit representation therefore, is to adopt a process theory. Essentially the same point applies to **O'Rourke**'s contention that a classicist could ground a vehicle theory in the distinction between explicitly represented rules (program) and explicitly represented information (data). This distinction is also grounded in computational relations, rather than intrinsic properties. Whether a sequence of tokens on a Turing machine's tape is an instruction or data is not determined by the tokens themselves; it is determined by their computational effects on the machine's read/write head. So we are thrown back to connectionism as the only plausible source of a vehicle theory. What this means, however, is that we must answer the question posed by **Dennett & Westbury**: If it is not their causal effects, what is it about the patterns of activity across the brain's neural networks that makes them conscious? It is time to put some content into our vehicle theory of consciousness. We begin this process in the next section.

R2. Representation

R2.1. Why link consciousness and representation? The connectionist theory of phenomenal experience proposed in the target article identifies consciousness with the brain's vehicles of explicit representation. Several commentators wonder about the general motivation for linking consciousness and representation in this way. **Perner & Dienes,** for example, observe that we spend a good deal of time reassessing the *dissociation* studies to make room for our proposal; but what we do not do, they think, is make out a strong case for *associating* representation and consciousness in the first place. This sentiment is echoed in various ways by **Ellis, O'Rourke,** and **Reeke.**

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

R2.2. Can vehicles be explicit representations? In the case of the connectionist vehicle theory that we defend, stable activation patterns across the brain's neural networks are presumed to be the relevant vehicles of explicit representation. A number of commentators have difficulties, however, both with our reliance on network activation patterns as vehicles in this respect, and with our reliance on the notion of explicit representation more generally. The general thrust of their objections is that any theory that identifies consciousness with either stable network activity or the explicit representation of information in the brain will inevitably incorporate elements of computational process: A "pure" vehicle theory of the sort we propose is not really a coherent option.

This charge is most straightforwardly developed by **Mac Aogáin** and **Wolters & Phaf.** They claim that network activation patterns cannot be treated as "free-standing" vehicles, either because they are the products of processes of relaxation (Wolters & Phaf) or because "there must be a process running in the background" to render them stable (Mac Aogáin). But we think both versions of this argument are invalid. Although it is true that stable patterns of activity are generated and sustained by flurries of activation passing that spread across networks, the suggested conclusion (that these patterns themselves must in part be understood *as* processes) does not follow. Network activation patterns are physical objects with intrinsic structural properties just as much as neurotransmitter molecules, neurons, and brains. That the latter entities need electrodynamic, biochemical, and neurophysiological processes to generate and support them in no way undermines their status as spatiotemporally extended objects in their own right.

Church develops the argument in a different way. Her worry is that the distinction we draw between explicit and nonexplicit representation, especially as we develop it in the connectionist context, is ill-begotten: "[N]either the notion of encoding in discrete objects nor the notion of active versus potentially active representation seems to help in specifying what is distinctive of [explicit] representation." However, Church's analysis is in one way incomplete, and in another way quite mistaken. It is mistaken in part because nowhere do we unpack the distinction between explicit and nonexplicit in terms of "active versus potentially active representations." Here Church seems to have improperly conflated "potentially explicit" with "potentially active." It is the potentially explicit information stored in a PDP network that governs its computational operations (target article, sect. 4.2). Thus, potentially explicit information is not merely potentially active; it is active every time a network is exposed to an input. More importantly, though, Church's analysis is incomplete because explicit representation, on our account, requires more than that information be coded by physically discrete objects. It also requires that the physical resources used to encode each item of information be distinct from those used to encode others. Activation pattern representations satisfy this requirement; no network pattern of activity represents more than one distinct content. We might say that they are *semantically* discrete. Connection weight representations, on the other hand, fail to do so, because they encode information in a superpositional fashion; each connection weight contributes to the storage of many, if not all, of the stable activation patterns (explicit representations) the network is capable of generating.

At this point we should introduce **Clapin**'s concerns about our representational taxonomy, as these specifically

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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

In the second prong of his assault **Clapin** insists that, contrary to our own analysis, activation patterns across PDP networks can and do participate in superpositional forms of information coding. He defends this claim by describing an imaginary three-layer PDP network that, because its input layer is divided into two subsets, is able to process two inputs simultaneously. He asserts that "in such a network the activation vector corresponding to the hidden units would superpositionally represent the two inputs." But here Clapin is using "superposition" merely to describe a process in which two (input) patterns are *combined* in some fashion to form a third (hidden layer) pattern. Superpositional *representation* is a quite different notion. As we have already noted, a computational device represents information in a superpositional fashion when the physical resources it employs to encode one item of information overlap with those used to encode others. It is standard practice in PDP modeling to suppose that the content of an activation pattern representation is fixed by the point it occupies in the network's "representational landscape." This landscape, which can be revealed by such numerical techniques as cluster analysis and principal components analysis, is constituted by the range of stable activation patterns that a trained-up network can generate in response to its full complement of possible inputs. Such a story about representational content seems to rule out superpositional coding in activation patterns, as it is impossible for an individual activation pattern to occupy, at one time, two or more different points in this representational landscape. (For a more detailed discussion of the issues raised in this paragraph, see Clapin & O'Brien 1998.)

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

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

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

It is fashionable in cognitive science and philosophy of mind to suppose that the mind's representational content must be unpacked in terms of causal transactions between the brain and the environment in which it is embedded. On this view content has very little to do with the intrinsic properties of the brain's representational vehicles, and everything to do with their causal relations (e.g., actual, counterfactual, or historical) with the world. It is for precisely this reason that **Lloyd** counsels us to dissociate phenomenal content from representational content. Because a vehicle theory holds that "states of consciousness are identical with states individuated by their intrinsic properties, rather than

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

by their functional context," he writes, it implies that conscious experiences "cannot carry information about an outside world, and so cannot represent an outside world (at least by most accounts of representation)." Lloyd's suggestion is that on the approach we are taking, an approach with which he is highly sympathetic, consciousness should not be viewed as a kind of representational vehicle, it should simply be understood as "a complex state."

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

R2.3. Connectionism and representational content. The standard analysis of representational content is, in large measure, a legacy of classical cognitive science. Given that digital computations inherit their semantic coherence from rules that are quite distinct from the structural properties of the symbols to which they apply, classicism appears to place few constraints on a theory of mental content. Thus the causal, functional, and teleofunctional theories that dominate the current literature are all, prima facie, compatible with the idea that mental representations are symbols. All of this changes, however, when we move across to connectionism. Given its foundation in the PDP computational framework, connectionism undermines the distinction between representational vehicles and the processes that act on them. A PDP system is not governed by rules that are distinct from the structural properties of its representational states. Instead, it computes by exploiting a *structural isomorphism* between its physical substrate and its target domain.

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

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

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

R3. Stability

R3.1. Why *stable* activation patterns? Lloyd, Mangan, and **Reeke,** all of whom express some sympathy for our proposal, wonder why we have restricted our account of consciousness to *stable* patterns of neural activity. This, they think, both prevents us from taking advantage of the "dynamic" properties of neural activity, and makes it impossible for our vehicle theory to explain the "flights" of consciousness – its fluidity and evanescence. With regard to the former, Reeke thinks it unfortunate that by focusing on stable patterns we have neglected the role that the complex reentrant interaction between neural patterns has on the emergence of phenomenal experience across large portions of the brain, and Mangan is aghast that we have needlessly ignored the contribution that stabilizing networks can make to conscious experience. With respect to the latter, Reeke questions how stability "even in suitably quantized chunks of time" can explain the smooth flow of experience, and Lloyd suggests that it is more likely that instability accounts for the unmemorability and nonreportability of certain kinds of conscious episodes.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

These worries are important. But as we observe in the target article (sect. 5.1), there is a straightforward reason for making stability such a central feature of our connectionist account. A vehicle theory identifies consciousness with the brain's explicit representation of information, and only stable patterns of activation are capable of encoding information in an explicit fashion in PDP systems. As we argued in the previous section, patterns of activation across PDP networks are contentful, by virtue of being structurally isomorphic with certain aspects of a target domain. But such structural isomorphs cannot be realized in PDP networks unless they achieve stable patterns of activity. This is because prior to stabilization, there are no objects physically present in these networks whose intrinsic structural properties can stand in this kind of relation to elements of the target domain. A connectionist vehicle theory must, therefore, identify phenomenal experience with stable patterns of activity across the brain's neural networks.

What is more, contrary to what these commentators claim, a vehicle theory that focuses on stable patterns can both exploit the dynamical properties of neural activity and explain the fluidity and evanescence of conscious experiences. Against **Reeke,** for example, the connectionist vehicle theory we are proposing does not neglect the complex interplay between large-scale patterns of activity across the brain. It is precisely this interplay, we think, that is responsible for the internetwork processes that bind phenomenal elements together to construct the unified cognitive subject (see sect. R7 below). Moreover, given the time scale at which network stabilizations can occur in the brain, it is not implausible to suppose that the "seamless" flow of experience is actually composed of quantized phenomenal elements, each equipped with its own distinct representational content. In this sense, the "flights" of consciousness that **Lloyd** (following James 1890) highlights, are plausibly reconceived as rapid sequences of "perchings," and it is their rapidity, not the absence of stability, that accounts for the unmemorability and nonreportability of certain kinds of conscious experiences.

R3.2. What is a stable activation pattern? We have just argued that a connectionist vehicle theorist is committed to identifying consciousness with stable patterns of activity across the brain's neural networks. According to a number of commentators, however, this just leads us into more hot water. For them, and here we have in mind **Cleeremans & Jiménez, Dennett & Westbury, Gilman, Lloyd, Pólya & Tarnay, Schröder,** and **Taylor,** our characterization of stability is problematic. In the target article we opt for a simple story according to which a neural network realizes a stable pattern when its constituent neurons are firing simultaneously at a constant rate (sect. 5.1). Several commentators observe that such a definition is incomplete, absent a relevant time scale. We accept this criticism. But this problem is not ours alone: Connectionist theorizing about cognition in general is deeply committed to stability.

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

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

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

R3.3. Stability in simulations and artificial networks. We will finish this section by examining a number of worries

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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

Dennett & Westbury disagree. They claim that the stability of a virtual machine is every bit as powerful as the stability of an actual machine. We think this is demonstrably false for the reasons we develop in the target article: There are vast temporal asymmetries between real PDP networks and their digital simulations. These temporal asymmetries arise because the structural properties (and hence causal powers) of a numerical array in a digital computer are quite different from those possessed by the web of active elements (however these might be physically realized) that make up a real PDP network. Most obviously, there are causal connections between the elements of a PDP network that simply do not exist between the variations in voltage that physically implement a numerical data structure. The manner in which activation patterns evolve in each case is thus quite different. In a real network, this evolution is dynamic: All the connected elements have their activity continuously modulated by the activity of others, until the network settles into a stable state. The activation pattern in a simulated network, by complete contrast, evolves through the operation of an algorithm that updates activation values individually.

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

Stated this baldly, we know that many readers (including **Perner & Dienes,** presumably) will balk at these implications of the vehicle theory we are defending. Do we really want to be committed to an account of consciousness that ascribes experiences to artificial networks? **Pólya & Tarnay** develop this worry in a particularly vivid way by noting that there is a sorites paradox looming here: Just how complex does a network pattern have to be before it is capable of conscious experiences? And **Gilman** simply asks: "Is a 28 cell network, stable in vitro, conscious? If so, of what?" Vehicle theorists are not completely without resources in this regard. They can appeal to the theory of representational content we briefly sketched above (sect. R2.2) to provide some minimal constraints on both the emergence of conscious experiences and the nature of their phenomenal properties. But there is a deeper issue here. The air of implausibility that surrounds our connectionist proposal at this juncture is one that envelops *all* materialist theories of consciousness. Whatever physical or functional property of the brain's neural networks one cares to name, one will always be vulnerable to these kinds of considerations. For example, process theorists who explain consciousness in terms of the rich and widespread processing relations enjoyed by a (relatively small) subset of the brain's representational vehicles have the problem of specifying just *how* rich and widespread these processing relations must be. Sorites considerations are, then, just another way of highlighting the much vaunted explanatory gap with which all materialists must contend. Toward the end of the target article we suggest how our connectionist proposal might close this gap (sect. 5.4). Not surprisingly, a number of commentators are not convinced. We address their concerns in the next section.

R4. Identity (or: Once more into the explanatory gap)

Suppose on some future day in the golden age of connectionist neuroscience we have compiled an exhaustive account of all the kinds of activation patterns that are generated in all the different neural networks in our heads. And suppose further that experimental research demonstrates that there is a precise one-to-one mapping between these kinds of activation patterns and specific types of phenomenal experience. Would this show that conscious experiences are *identical* to stable patterns of network activity in the brain? **Ellis, Kurthen, Newton, van Heuveln & Dietrich,** and **Velmans** think not. Our connectionist vehicle theory, they charge, whatever its virtues as an account of the

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

neural *correlates* of consciousness, does not have the resources to bridge the explanatory gap.

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

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

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

Suppose, then, that in addition to showing that there is a one-to-one mapping between activation patterns and types of phenomenal experience, our future neuroscience also reveals that in every case there is a structural isomorphism between these patterns and the properties of the world represented in the corresponding phenomenal experiences. Would this be enough to close the door on consciousness? **Velmans** thinks not and develops what is perhaps the most familiar line of reasoning that maintains a distance between our hypothesis about the neurocomputational substrate of consciousness and its phenomenal properties: "One might . . . know everything there is to know about the 'shape' and 'dimensionality' of a given neural activation space," he writes, "and still know nothing about what it is like to have the corresponding experience." Indeed, anticipating our earlier discussion of artificial networks, Velmans, with a Nagelian twist (1974), asks us to suppose that we arrange "a net to operate in a nonhuman configuration, with an 'activation space shape' which is quite unlike that of the five main, human, sensory modalities." According to Velmans, we cannot know what such an artificial network would experience. And "if we can know the 'shape' of the space very precisely and still do not know what it is like to have the experience, then having a particular activation [pattern] cannot be *all there is to having an experience.*"

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

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

Regardless of whatever else might be necessary, knowledge implicates representation. On the theory of representation that we are considering, this requires the presence of structural isomorphisms between patterns of neural activity and certain aspects of the thing known. Consider what it takes to have knowledge about phenomenal experiences (as opposed to knowledge about aspects of the world derived from phenomenal experiences). Because phenomenal experiences are activation patterns across neural networks, knowing about them requires one to generate structural isomorphs of these patterns. But now something very interesting happens. To have *exhaustive* knowledge about a phenomenal experience, on this analysis, one must gener-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

ate a neural activation pattern that is exactly structurally isomorphic with the activation pattern one seeks to know – a pattern that is in turn isomorphic with some aspect of the world. Transitivity of structural isomorphism thus dictates that to have complete knowledge about a phenomenal experience, one must *reproduce* this very experience in one's head.

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

Part II: Empirical plausibility

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

R5. General empirical concerns

R5.1. Consciousness is limited, but the unconscious is vast. A very general worry raised by **O'Rourke** is that our picture of unconscious mental activity is seriously awry, because it inverts the usual picture, championed by Baars (1998; 1994), of the relationship between the conscious and the unconscious: the former is limited, serial, and slow; the latter vast, parallel, and speedy. O'Rourke thinks that by excluding the possibility of unconscious explicit representations, we seriously limit the role of the unconscious. Consequently, we are not in a position to offer serious candidate explanations for even simple cognitive phenomena such as recall.

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

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

Incidentally, this last point is something that **Velmans** disputes. He rightly observes that information unconsciously represented in memory is constantly "causally active in determining our expectations and interactions with the world," and thinks that it must, therefore, be explicitly encoded. It is hard to see how such a view could be sustained. First, it is contrary to the orthodox understanding of both digital and PDP systems, whereby tacitly (and hence, nonexplicitly) represented information plays a pivotal computational role. Second, and perhaps more importantly, it is precisely because classicism is committed to a vast amount of unconscious, causally discrete information processing that the infamous "frame problem" is so acute for this approach to cognition. The connectionist vehicle theory, with its promise of a causally holistic unconscious, appears ideally placed to provide a more realistic solution.

R5.2. Are there not many stable activation patterns? It is at this point that a very common objection comes to the fore. Are there not simply too many stable patterns of activation in the brain for us to pursue seriously a connectionist vehicle theory of consciousness? This objection is raised, in one form or another, by **Cleeremans & Jiménez, Gilman, Mangan, Perner & Dienes,** and **Van Gulick.**

Mangan claims that our account is "at odds with virtually all existing PDP models of neural activity." He takes it to be a fundamental feature of connectionist theorizing that there are lots of relaxed PDP networks in the brain that do not generate any conscious experience. This is a curious claim. Connectionist theorizing about phenomenal consciousness is in its infancy. Mangan gives the impression that we are swimming against a great tide of connectionist thinking in this area, but we think it is premature to start

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

looking for a consensus. Apart from some suggestive comments in McClelland and Rumelhart (1986), Smolensky (1988), and Mangan (1993b), and the pioneering work of Lloyd (1991; 1995a; 1996), there has not really been much work on the foundations of a distinctively connectionist approach to consciousness. The ultimate shape of such a theory is still very much up for grabs. It is in the spirit of exploring largely uncharted territory that we make our suggestions.

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

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

When it comes to perception, and in particular to the early processing of visual, auditory, and speech signals, classical accounts presuppose a great deal of unconscious, explicitly represented information. Perceptual processing is assumed to be hierarchical in nature, beginning with a first stage of representations that are transduced from environmental input, transformations of which lead to further interlevels of explicit representation.[3] The contents of sensory consciousness correspond with some privileged stage in the processing of input,[4] but the vast majority of the explicit representations generated during input processing are taken to be unconscious. **Gilman** and **Van Gulick** seem to be persuaded by these default (process model) assumptions. However, we do not think they are compulsory for a connectionist. The analysis we presented in our target article (sect. 5.2) suggests that phenomenal consciousness is exceedingly rich, far richer than classically-inspired theorists are generally willing to acknowledge. Moment-by-moment perceptual experience, in particular, embraces a multitude of object features and properties, at many levels of detail. In other words, there is actually a great deal of phenomenology to play with when framing connectionist theories of perception. Consequently, connectionists are in a position to advance PDP models of, say, vision, that posit a raft of stable activation patterns, without in any way undermining the connectionist vehicle theory of consciousness.

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

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

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

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

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

R5.3. Noncontrastive analyses. If the worries discussed in the previous section are less than devastating, they do at least flag the crucial difficulty for a defender of the connectionist vehicle theory of consciousness: How, in light of existing studies, does one justify the identification of phenomenal experience with the explicit representation of information in the brain? This problem has two parts, which line up with Dulany's useful distinction between *contrastive* and *noncontrastive* analyses (1991, pp. 107–11).

A contrastive analysis is one that makes differential predictions explicitly designed to test for the presence of unconscious, explicit representations. The dissociation studies (target article, sect. 2) are of this type. The first problem facing a vehicle theorist is to account for the weight of relatively direct evidence, provided by contrastive analyses,

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

for the dissociation of phenomenal experience and explicit representation. We revisit this issue in the next section.

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

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

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

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

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

However, the central point here is well taken. Where noncontrastive analyses have been successfully applied to cognitive phenomena we surely have some reason to take seriously any unconscious, explicitly represented information to which they appeal. That said, it is difficult to know how to assess this objection. Until recently, theorizing in cognitive science has been dominated by the classical conception of cognition (this is certainly true of Marr's theory of vision). We have argued that classicists are committed to the existence of explicit representations whose contents are not conscious (target article, sect. 3.2), so it is no surprise that such representations are legion in current theorizing. An inference to the best explanation is always vulnerable to the emergence of a rival theory with comparable simplicity and explanatory scope. With the advent of connectionism a whole new *class* of explanations is coming onto the scene, and it is no longer safe to assume that classically inspired theories will retain their favored status.

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

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

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

R6. The dissociation studies revisited

Not surprisingly, a number of commentators have taken us to task over our treatment of the dissociation studies: that large body of work that is widely interpreted as having established the existence of unconscious, explicitly represented information. **Perner & Dienes** claim that our review of the literature is "selective and dated," a sentiment echoed by **Velmans.** These commentators argue that the evidence for dissociation between conscious experience and explicit mental representation is more compelling than we allow, and point to specific studies we neglected. Velmans makes the additional point that we are burdened with demonstrating that *all* of the contrastive analyses conducted to date are flawed, thus "even *one* good example of preconscious or unconscious semantic processing would be

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

troublesome for [the connectionist vehicle theory of consciousness]."

These objections are well taken. However, in our defense, we took our role in the target article to be one of establishing that a connectionist vehicle theory of consciousness is not *conclusively* ruled out by existing studies. In this connection, it is interesting to note the rather different appraisal of the literature offered by **Dulany,** who describes the dissociation studies as being subject to "decades of conceptual confusion and methodological bias." He continues: "Suffice it to say now that if claims for the power of a cognitive unconscious were correct, the experimental effects would be too strong and replicable for these literatures even to be controversial. No one can claim that." This level of disagreement among cognitive psychologists certainly suggests that it is legitimate to pursue a vehicle theory, while suspending judgement regarding the dissociation studies. Nevertheless, **Velmans** is right about the burden we must ultimately shoulder. The onus is on a vehicle theorist to show that each of the contrastive paradigms is flawed in some way, or is open to reinterpretation in the light of connectionism. So we will make some further remarks about implicit learning, take a second look at the phenomenon of priming, and finish with a discussion of the more recent literature on blindsight.

R6.1. Implicit learning. In our discussion of implicit learning we relied heavily on the work of Shanks and St. John (1994), who characterize this phenomenon as the induction of unconscious rules from a set of rule-governed training stimuli. **Perner & Dienes**'s commentary suggests, and a survey of the more recent literature confirms, that this is not the only way theorists are apt to characterize implicit learning (see, e.g., Cleeremans 1997; Dienes & Berry 1997; Perruchet & Gallego 1997; Perruchet & Vinter 1998). A less contentious way of defining implicit learning, inspired by Perruchet and Gallego 1997 (p. 124), is: "An adaptive process whereby subjects become sensitive to the structural features of some stimulus domain without consciously deploying learning strategies to do so."[8] This definition captures what occurs in the standard experimental paradigms designed to investigate implicit learning: artificial grammar learning, instrumental learning, serial reaction time learning, and so on. However, it is sufficiently generic to cover aspects of first and second language learning, the acquisition of reading and writing, and adaptation to physical and social constraints. The essential contrast is with cases where a subject is aware that learning is taking place, and deploys various strategies to facilitate the process (for example, consciously forming and testing hypotheses about the stimulus domain).

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

Denying a particular explanation of implicit learning does not amount to denying the existence of the phenomenon, however. The issue for us now becomes: Is there an explanation of implicit learning (as defined above) that is compatible with the connectionist vehicle theory of consciousness? We think the answer is yes. The shape of such an explanation is suggested by **Perner & Dienes,** and finds its most detailed elaboration (minus the connectionist gloss) in the work of Perruchet and Gallego (1997) and Perruchet and Vinter (1998), who propose a *subjective unit-formation* account of implicit learning, in which "intrinsically unconscious" associative mechanisms generate increasingly appropriate parsings of the stimuli in some domain. First exposure to stimulus material brings on line specialized mechanisms – which may be innate, or the product of earlier learning – that parse stimuli into a set of small disjunctive units. These subjective units comprise our experience of the domain. They are selected and modified by subsequent training, and can go on to form the basis of higher-level subjective units (which fits with the way training both focuses and enriches our experience; see Perruchet & Vinter 1998, pp. 502–503 for a summary of this account).

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

R6.2. Subliminal priming. Several commentators raise concerns about our treatment of subliminal priming. **Perner & Dienes** claim that our critique of Marcel's (1983) experiments "rehashes old arguments already dealt with [by Marcel]," and that we have failed to mention more recent studies. **Velmans** draws attention to the work of Groeger (1984) who found evidence of semantic priming, on a subsequent word selection test, by subliminally presented words. He also refers to studies that appear to demonstrate that attention to a spoken word preconsciously activates all of its possible meanings for a short period (around 250 msec); "Depending on context, one meaning is selected and the subsequent entry of the word into consciousness is accompanied by inhibition (or deactivation) of inappropriate meanings." **Zorzi & Umiltà** remind us of the priming studies that have been done with subjects suffering from unilateral neglect (Berti & Rizzolatti 1992; Làdavas et al. 1993). For example, objects presented to the neglected hemifield of subjects with severe neglect appear to facilitate (i.e., speed up) responses to semantically related objects presented to the contralateral field.

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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

The reasoning behind the various priming studies seems to be that, for a subliminal stimulus to affect ongoing cognitive processing, an explicit representation of some sort has to be generated and manipulated in the cognitive system. If this happens, subliminal priming provides clear evidence for the existence of explicit representations with unconscious contents. But there is some question as to whether subliminal priming, *when interpreted from within the connectionist camp,* unequivocally leads to this conclusion. In particular, one should always be mindful of the *style* of computation employed by connectionist systems. A recurrent network, for example, when initially exposed to an input may oscillate quite dramatically as activation circulates around the network, and hence may take some time to relax into a stable pattern of activation (though here, of course, we are talking in terms of milliseconds). Just prior to stabilization, however, as these oscillations abate, the network is likely to converge on some small region of activation space. It is this feature of PDP processing that provides the leverage for a connectionist explanation of subliminal priming.

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

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

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

R6.3. Blindsight. A lot has happened in blindsight research of late, as **Perner & Dienes** and **Kentridge** rightly point out. The stray light hypothesis, offered by Campion (1983) as a possible explanation of blindsight, appears to have been taken on board and controlled for (see, e.g., Milner & Goodale 1995, pp. 72–73, and Shallice 1997, p. 258, for discussion). The spared cortex hypothesis also looks less plausible in light of recent work (Kentridge et al. 1997). The evidence is now fairly conclusive that a range of visually guided behaviors can occur without striate mediation, and, as Kentridge points out, that both subcortical and cortical

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

structures are implicated in these behaviors (see Tobée 1996, pp. 71–74, for a brief discussion).

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

What of these recent studies? All of them have involved testing a single subject, GY, whose left visual cortex is badly damaged, such that he is clinically blind in his right hemifield. Weiskrantz et al. (1995) tested GY on a motion discrimination task using a two-alternative forced-response procedure. He responded by key press, using one of two keys to indicate the direction of motion of a moving spot presented in his blind hemifield, and one of two keys to indicate, respectively, no awareness, or some awareness. Awareness was tested after every trial, in blocks of 50 or 100 trials. On those trials where he signalled no awareness GY was still required to guess a direction of motion. Striking results were obtained. Although GY often reported visual experiences of some kind, in those instances where he reported no awareness whatever he achieved as high as 90% accuracy (across a block of trials) for direction of motion judgments.

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

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

Marcel's (1993) study of GY adds weight to the conviction that the relationship between performance and experience in GY is still to be fully explored. GY's task was simply to respond to a spot of bright light in his blind hemifield. In one experiment he was asked to respond using three different modalities at once: blinking, button pressing, and verbalizing. There were frequent dissociations between the modalities; GY would, for example, report "yes" with his eye, but not with his finger. In a second experiment the various modalities were tested separately. Marcel ran 8 blocks of 40 trials for each, in which GY was asked to "'respond as soon as you have an impression of whether a light came on'" (the "report" condition; 1993, p. 170). In a second condition (the "guessing" condition) two additional blocks of trials were run in which GY was instructed to "*guess* whether a light had come on, irrespective of any conscious sensation" (p. 171). Marcel found that in the first two blocks of trials the success rates were similar across modalities, although performance was marginally better for the blink response. However, there was a marked training effect for the blink and button press. Before training the percentages of hits (correct positive responses) for the blink and button presses were 67.5% and 62.5%, respectively. After training, this figure had jumped to 70% for the button press, and 85% for the blink response! Verbal response did not improve with training, and performance stayed around 60% hits (1993, p. 172). Most striking, in the guessing condition Marcel found that GY was able to achieve on the order of 80% hits (87.5% for the blink response), *irrespective of response modality.* In other words, GY's performance using button press and verbal assent when merely guessing, approached the results achieved using blink response when attending to visual experience.

How does one interpret these results? To begin with, it is reasonable to suppose that the improvement in blink response performance is reflected in alterations to structures in the dorsal stream, where blink responses are known to be controlled (see Milner & Goodale 1993). What is significant is that performance across all response modalities, under the instruction to *guess,* is no better than the trained-up performance for blinking, under the instruction to respond to a visual experience. The most parsimonious reading of these results is that the very phenomenology driving the blink response is also responsible for performance in the guessing condition. That is, guessing is actually guided by

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

some sort of phenomenology, but phenomenology that has only an indirect effect on button presses and verbal responses. One might think, in fact, that the phenomenology in this case is visuomotor, given the seemingly crucial role of the dorsal stream in these results. Vision (1998) remarks that "[p]erhaps . . . GY did not feel compelled to perceive the light on the guessing trials, but only to use whatever feelings he had available for answers (even if they were based on feelings acquired from premotor muscular preparation)" (p. 151).

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

R7. Subject unity, agency, and introspection

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

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

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

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

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

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

As for the coordination of control, a proper answer to this question would essentially involve us in providing a complete connectionist account of cognition. Such an account would show, in detail, how perceptual, cognitive, and motor processes are integrated, and would at every point indicate the role of phenomenal experience in this integration. It would also address many of the issues that rightly concern **Dulany:** the nature of propositional contents, delib-

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

erative thought, metacognitive awareness, and so on. We have done some preliminary work on these issues (Opie 1998, Ch. 7), but clearly much remains to be done.

McDermott expresses concern that on our account the link between introspection and consciousness is not necessary; that we allow for "the bizarre possibility that most of our conscious experiences are not accessible to introspection." If we follow McDermott, and treat introspection as the process whereby one conscious state becomes the object of another, then introspection is surely a common feature of human thought. However, this is no surprise on the connectionist vehicle theory of consciousness. Given the massive connectivity of the brain, and the role of stable activation patterns in internetwork processing, one would expect almost every phenomenal experience to be available, in principle, to introspection and verbal report. It is only under pathological, or degraded input conditions, that one would anticipate any deviation from this.

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

Similar confusions abound in the literature concerning the relationship between consciousness and attention. Some theorists conflate the two, and so take a theory of attention to be none other than a theory of phenomenal consciousness. Elsewhere we argue that this conflation is fundamentally mistaken (O'Brien & Opie 1998). Consciousness incorporates both a central focus and a rich polymodal periphery. Theorists often neglect the periphery, but it is important nevertheless (it can save your life when crossing the road). Both focus and periphery are parts of one's instantaneous experience, so the attended/unattended distinction does not line up with the conscious/unconscious distinction, **Coltheart**'s suggestions notwithstanding. The standard view, as we understand it, is that cognition depends on an enormous number of mental representations whose contents are not merely peripheral, but completely absent from consciousness. Incidentally, we cannot agree with Coltheart that it is only possible to attend to one thing at a time. One can, for example, hold a conversation and deal cards at the same time. Support for this intuition comes from recent studies that have led to the proposal of a multiple resource theory of attention (see Anderson 1995, pp. 103–104 for discussion).

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

> a procession, where half the people taking part are deaf and the other half are blind. . . There will be many visual experiences and at the same time many auditory ones. But none of this generates the unified experience of both seeing and hearing the procession. (1988, pp. 54–55)

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

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

If this is the problem, then our response to **Dulany** and **Carlson** is the beginning of a solution – one that relies on internetwork processing to maintain the coherence of experience, and to generate the multiple abstract contents that constitute a sense of unity. Our suggestions about narrative and point of view may be seen as contributions to the task of analyzing this "sense of unity" (in its diachronic and synchronic forms, respectively).

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

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

NOTES

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

2. Two theorists who have kept the torch of structural isomorphism burning over the years are Palmer (1978) and Shepard (Shepard & Chipman 1970; Shepard & Metzler 1971). But more recently, Blachowicz (1997), Cummins (1996), Edelman (1998), Files (1996), Gardenfors (1996), and Swoyer (1991) have all explored, though in different ways, the idea that relations of isomorphism might ground the content of the brain's representational vehicles. For a general discussion of these issues, see O'Brien (forthcoming).

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

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

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

6. See, for example, Arbib and Hanson (1987) and Lehky and Sejnowski (1990).

7. This is probably not the way to tackle delayed recall. One promising approach is suggested by the work of Smith and

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Blankenship (1989; 1991). Their examination of unconscious incubation, which is of a piece with delayed recall, suggests that incubation amounts to no more than the gradual decay of a memory block. This explanation is essentially noncognitive, in that it does not appeal to processes defined over information-bearing states. For other explanations of incubation along these lines see Boden (1990), pp. 244–45, and Perkins (1981).

8. See also Cleeremans (1997), section 3.

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

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

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

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

13. That there is more to say about GY's phenomenology is suggested by a remark reported in Zeki and ffytche (1998). At one point, when presented with a low contrast stimulus, GY "spontaneously remarked that the awareness score here should be 'minus one or minus two'" (p. 30). Zeki and ffytche take this to imply that there might be "degrees of unawareness" for GY. This seems a very odd conclusion, and the idea of "degrees of unawareness" hardly seems coherent. The more natural conclusion, one might think, is that even when GY reports a 1 (corresponding to no awareness) on the 4-level scale, he actually does have some kind of phenomenology – phenomenology compared to which a genuine "no awareness" states looks like a "minus one or two."

References

The letters "a" and "r" before authors' initials refer to target and response article references, respectively.

Akins, K. (1996) Lost the plot? Reconstructing Dennett's multiple drafts theory of consciousness. *Mind and Language* 11:1–43. [aGO]

Anderson, J. R. (1995) *Cognitive psychology and its implications,* 4th edition. Freeman. [rGO]

Arbib, M. & Hanson, A. (1987) *Vision, brain, and cooperative computation.* MIT Press. [rGO]

Aurell, C. (1989) Man's triune conscious mind. *Perceptual and Motor Skills* 68:747–54. [RE]

Azzopardi, P. & Cowey, A. (1997) Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences USA* 94:14190–94. [RWK]

Baars, B. J. (1988) *A cognitive theory of consciousness.* Cambridge University Press. [arGO]

(1994) A thoroughly empirical approach to consciousness. *Psyche* 1(6). **http://psyche.cs.monash.edu.au/v2/psyche-1–6-baars.html**. [rGO]

(1996) *In the theater of consciousness: The workspace of the mind.* Oxford University Press. [DED, GW]

(1997) In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4:292–309. [JO]

Baddeley, A. (1986) *Working memory.* Oxford University Press. [JGT, GW]

Barbur, J. L., Harlow, A. J. & Plant, G. (1994) Insights into the different exploits of colour in the visual cortex. *Proceedings of the Royal Society of London, Series B* 258:327–34. [RWK]

Bechtel, W. (1988a) Connectionism and rules and representation systems: Are they compatible? *Philosophical Psychology* 1:1–15. [aGO]

(1988b) Connectionism and interlevel relations. *Behavioral and Brain Sciences* 11:24–25. [aGO]

Bechtel, W. & Abrahamsen, A. (1991) *Connectionism and the mind.* Blackwell. [aGO]

Becker, S., Moscovitch, M., Behrmann, M. & Joordens, S. (1997) Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23:1059–82. [AC]

Berti, A. & Rizzolatti, G. (1992) Visual processing without awareness: Evidence from unilateral neglect. *Journal of Cognitive Neuroscience* 4:345–51. [rGO, MZ]

Bickhard, M. H. & Terveen, L. (1995) *Foundational issues in artificial intelligence and cognitive science: Impasse and solution.* North Holland. [TP]

Blachowicz, J. (1997) Analog representation beyond mental imagery. *Journal of Philosophy* 94:55–84. [rGO]

Block, N. (1993) Book review of Dennett's *Consciousness explained. Journal of Philosophy* 90:181–93. [aGO]

(1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18:227–87. [arGO, JO]

Boden, M. (1990) *The creative mind.* Abacus. [rGO]

Bower, G. H. (1975) Cognitive psychology: An introduction. In: *Handbook of learning and cognitive processes* (vol. 1), ed. W. K. Estes. Erlbaum. [DED]

Cam, P. (1984) Consciousness and content-formation. *Inquiry* 27:381–97. [aGO]

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. [RWK, arGO]

Carlson, R. A. (1992) Starting with consciousness. *American Journal of Psychology* 105:598–604. [RAC]

(1997) *Experienced cognition.* Erlbaum. [RAC]

Carlson, R. A. & Dulany, D. E. (1988) Diagnostic reasoning with circumstantial evidence. *Cognitive Psychology* 20:463–92. [DED]

Carruthers, P. (1996) *Language thought and consciousness. An essay in philosophical psychology.* Cambridge University Press. [JP]

Chalmers, D. J. (1995) Facing up to the problem of consciousness. *Journal of Consciousness Studies* 2:200–19. [aGO]

(1996) *The conscious mind: In search of a fundamental theory.* Oxford University Press. [AC, aGO]

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

Chomsky, N. (1980) Rules and representations. *Behavioral and Brain Sciences* 3:1–62. [aGO]

Churchland, P. M. (1990) Knowing qualia: A reply to Jackson. In: *A neurocomputational perspective.* MIT Press. [rGO]

(1995) *The engine of reason, the seat of the soul.* MIT Press. [arGO, RVG, BvH, GW]

Churchland, P. S. & Sejnowski, T. (1992) *The computational brain.* MIT Press. [arGO, TP]

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

Clark, A. (1989) *Microcognition: Philosophy, cognitive science, and parallel distributed processing.* MIT Press. [aGO]

(1993a) *Associative engines: Connectionism, concepts and representational change.* MIT Press/Bradford Books. [HC, aGO]

(1993b) *Sensory qualities.* Oxford. [DL]

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. [AC, aGO]

Clark, A. & Thornton, C. (1997) Trading spaces: Computation, representation and the limits of uninformed learning. *Behavioral and Brain Sciences* 20:57–90. [aGO]

Cleeremans, A. (1997) Principles for implicit learning. In: *How implicit is implicit learning?,* ed. D. C. Berry. Oxford University Press. [AC, rGO]

Cleeremans, A. & Jiménez, L. (submitted) Implicit cognition with the symbolic metaphor of mind: Theoretical and methodological issues. [AC]

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

Collins, A. M. & Loftus, E. F. (1975) A spreading activation theory of semantic processing. *Psychological Review* 82:407–28. [MZ]

Copeland, B. J. (1997) The broad conception of computation. *American Behavioral Scientist* 40:690–716. [aGO]

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

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

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

Cottrell, A. (1995) *Tertium datur?* Reflections on Owen Flanagan's *Consciousness reconsidered. Philosophical Psychology* 8:85–103. [aGO]

Cowey, A. (1996) Visual awareness: Still at sea with seeing? *Current Biology* 6(1):45–47. [rGO]

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

Crick, F. & Koch, C. (1995) Are we aware of neural activity in primary visual cortex? *Nature* 375:121–23. [DG]

(1998) Consciousness and neuroscience. *Cerebral Cortex* 8:97–107. [DG]

Cummins, R. (1986) Inexplicit representation. In: *The representation of knowledge and belief,* ed. M. Brand & R. Harnish. University of Arizona Press. [HC, aGO]

(1989) *Meaning and mental representation.* MIT Press. [rGO]

(1996) *Representations, targets, and attitudes.* MIT Press. [arGO]

Cummins, R. & Schwarz, G. (1991) Connectionism, computation, and cognition. In: *Connectionism and the philosophy of mind,* ed. T. Horgan & J. Tienson. Kluwer. [GO]

Cussins, A. (1990) The connectionist construction of concepts. In: *The philosophy of artificial intelligence,* ed. M. Boden. Oxford University Press. [aGO]

Damasio, A. R. (1994) *Descartes' error.* Putnam. [NN]

Dennett, D. C. (1981) Three kinds of intentional psychology. In: *Time, reduction and reality,* ed. R. Healy. Cambridge University Press. (Reprinted in Dennett, D. C. (1987) *The intentional stance.* MIT Press). [RVG]

(1982) Styles of mental representation. *Proceedings of the Aristotelian Society* (New Series) 83:213–26. [JC, DCD, DED, arGO, TP, MSCT] Reprinted in *The Intentional Stance.* MIT Press. [HC]

(1984) Cognitive wheels: The frame problem of A1. In: *Minds, machines and evolution,* ed. C. Hookway. Cambridge University Press. [aGO]

(1987) *The intentional stance*. MIT Press/A Bradford Book. [DCD]

(1991) *Consciousness explained.* Little, Brown. [DCD, MK, arGO]

(1993) The message is : There is no *m*edium. *Philosophy and Phenomenological Research* 53:919–31. [arGO]

(1998) The myth of double of transduction. In: *Toward a science of consciousness, II,* ed. S. R. Hameroff, A. W. Kaszniak & A. C. Scott. MIT Press/Bradford Books. [DCD]

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. [TP]

Dienes, Z. & Berry, D. (1997) Implicit learning: Below the subjective threshold. *Psychonomic Bulletin and Review* 4:3–23. [rGO, JP]

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* 17:875–78. [aGO]

Dietrich, E. (1989) Semantics and the computational paradigm in cognitive psychology. *Synthese* 79:119–41. [aGO]

Dixon, N. F. (1971) *Subliminal perception: The nature of a controversy*. McGraw-Hill. [MV]

(1981) *Preconscious processing.* Wiley. [MV]

Dretske, F. (1993) Conscious experience. *Mind* 102:263–83. [MK, aGO]

(1995) *Naturalizing the mind.* MIT Press. [aGO]

Dulany, D. E. (1968) Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In: *Verbal behavior and general behavior theory,* ed. T. Dixon & D. Horton. Prentice-Hall. [DED]

(1991) Conscious representation and thought systems. In: *Advances in social cognition IV,* ed. R. S. Wyer, Jr. & T. K. Srull. Erlbaum. [DED, arGO]

(1997) Consciousness in the explicit (deliberative) and implicit (evocative). In: *Scientific approaches to consciousness,* ed. J. Cohen & J. Schooler. Erlbaum. [DED, aGO]

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. [aGO]

Edelman, G. M. (1987) *Neural Darwinism: The theory of neuronal group selection.* Basic Books. [GNR]

(1989) *The remembered present: A biological theory of consciousness.* Basic Books. [aGO]

Edelman, S. (1998) Representation is the representation of similarities. *Behavioral and Brain Sciences* 21(4):449–98. [rGO]

Ellis, R. (1995) *Questioning consciousness: The interplay of imagery, cognition and emotion in the human brain.* John Benjamins. [RE]

Ellis, R. D. & Newton, N. (1998) Three paradoxes of phenomenal consciousness. *Journal of Consciousness Studies* 5(4):419–42. [NN]

Elman, J. L. (1990) Finding structure in time. *Cognitive Science* 14:179–211. [AV]

Ericsson, K. A. & Simon, H. A. (1993) *Protocol analysis.* MIT Press. [DL]

Farah, M. J. (1994a) Visual perception and visual awareness after brain damage: A tutorial overview. In: *Attention and performance XV: Conscious and nonconscious information processing,* ed. C. Umiltà & M. Moscovitch. MIT Press/Bradford Books. [MSCT, MZ]

(1994b) Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences* 17:43–104. [MSCT]

Farah, M. J., O'Reilly, R. C. & Vecera, S. P. (1993) Dissociated overt and covert recognition as an emergent property of a lesioned neural network. *Psychological Review* 100:571–88. [MSCT]

Field, H. (1978) Mental representation. *Erkenntenis* 13:9–61. [aGO]

Files, C. (1996) Goodman's rejection of resemblance. *British Journal of Aesthetics* 36:398–412. [rGO]

Flanagan, O. (1992) *Consciousness reconsidered.* MIT Press. [arGO]

Flohr, H. (1991) Brain processes and phenomenal consciousness: A new and specific hypothesis. *Theory and Psychology* 1:245–62. [RVG]

Fodor, J. A. (1975) *The language of thought.* MIT Press. [aGO]

(1981) *Representations.* MIT Press. [DL, aGO]

(1983) *The modularity of mind.* MIT Press. [arGO]

(1987) *Psychosemantics.* MIT Press. [aGO]

(1990) *A theory of content.* MIT Press. [MK]

Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71. [aGO]

Fuster, J. (1989) *The prefrontal cortex.* Raven Press. [JGT]

Gardenfors, P. (1996) Mental representation, conceptual spaces and metaphors. *Synthese* 106:21–47. [arGO]

Garfield, J. (1997) Mentalese not spoken here: Computation, cognition and causation. *Philosophical Psychology* 10:413–35. [JO]

Gawne, T., Kjaer, T. & Richmond, B. (1996) Latency: Another potential code for feature binding in striate cortex. *Journal of Neurophysiology* 76(2):1356–60. [DG]

Gibson, J. J. (1979) *The ecological approach to visual perception.* Houghton-Mifflin. [RAC]

Glover, J. (1988) *I: The philosophy and psychology of personal identity.* Penguin Press. [rGO]

Goldman-Rakic, P. (1996) Memory recording experiences in cells and circuits-diversity in memory research. *Proceedings of the National Academy of Sciences USA* 93:13435–37. [JGT]

Green, C. (1998) Are connectionist models theories of cognition? *Psycoloquy* 9(4):1–11. [DED]

Greenwald, A. G. (1992) New look 3: Unconscious cognition reclaimed. *American Psychologist* 47:766–79. [JP, GW]

Groeger, J. A. (1984) Evidence of unconscious semantic processing from a forced error situation. *British Journal of Psychology* 75:305–14. [rGO, MV]

Hardin, C. L. (1988) *Color for philosophers.* Hackett. [aGO]

Harman, G. (1973) *Thought.* Princeton University Press. [aGO]

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

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

(1985) *Artificial intelligence: The very idea.* MIT Press. [aGO]

Hebb, D. O. (1949) *Organization of behavior.* Wiley. [DCD]

Heller, J., Hertz, J., Kjaer, T. & Richmond, B. (1995) Information flow and temporal coding in primate pattern vision. *Journal of Computational Neuroscience* 2:175–93. [DG]

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

Heywood, C. A., Kentridge, R. W. & Cowey, A. (1998) Cortical colour blindness is not 'blindsight for colour.' *Consciousness and Cognition* 7:410–23. [RWK]

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. [arGO, MV]

Hopcroft, J. E. & Ullman, J. D. (1979) *Introduction to automata theory, languages and computation.* Addison Wesley. [aGO]

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA* 79:2554–58. [GNR]

Horgan, T. & Tienson, J. (1989) Representations without rules. *Philosophical Topics* 27:147–74. [aGO]

(1996) *Connectionism and the philosophy of psychology.* MIT Press. [aGO]

Hummel, J. E. & Holyoak, K. J. (1997) Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104:427–66. [DED]

Hurley, S. (1993) Unity and objectivity. *Proceedings of the British Academy* 83:49–77. [rGO]

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Jackendoff, R. (1987) *Consciousness and the computational mind.* MIT Press. [DED, arGO]
Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly* 32:127–36. [rGO]
James, W. (1890) *The principles of psychology.* Holt. [DL, arGO]
Johnson-Laird, P. N. (1988) *The computer and the mind: An introduction to cognitive science.* Fontana Press. [aGO]
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. [aGO]
Johnston, W. A. & Wilson, J. (1980) Perceptual processing of nontargets in an attention task. *Memory and Cognition* 8:372–77. [aGO]
Keijzer, F. A. & Bem, S. (1996) Behavioral systems interpreted as autonomous agents and as coupled dynamics systems. *Philosophical Psychology* 9:323–46. [EMA]
Kentridge, R. W., Heywood, C. A. & Weiskrantz, L. (1997) Residual vision in multiple retinal locations within a scotoma: Implications for blindsight. *Journal of Cognitive Neuroscience* 9:191–202. [RWK, rGO]
(in press) Effects of temporal cueing on residual visual discrimination in blindsight. *Neuropsychologia.* [RWK]
Kihlstrom, J. F. (1987) The cognitive unconscious. *Science* 237:1445–52. [GW]
(1996) Perception without awareness of what is perceived, learning without awareness of what is learned. In: *The science of consciousness: Psychological, neuropsychological, and clinical reviews,* ed. M. Velmans. Routledge. [MV]
King, S. M., Azzopardi, P., Cowey, A., Oxbury, J. & Oxbury, S. (1996) The role of light scatter in the residual visual sensitivity of patients with complete cerebral hemispherectomy. *Visual Neuroscience* 13:1–13. [RWK]
Kinsbourne, M. (1988) Integrated field theory of consciousness. In: *Consciousness in contemporary science,* ed. A. Marcel & E. Bisiach. Clarendon Press. [aGO, RVG]
(1995) Models of consciousness: Serial or parallel in the brain? In: *The cognitive neurosciences,* ed. M. Gazzaniga. MIT Press. [aGO]
Kirsh, D. (1990) When is information explicitly represented? In: *Information, language, and cognition,* ed. P. P. Hanson. University of British Columbia Press. [HC, rGO, JS]
Kleene, S. C. (1967) *Mathematical logic.* Wiley. [aGO]
Kurthen, M. (1995) On the prospects of a naturalistic theory of phenomenal consciousness. In: *Conscious experience,* ed. T. Metzinger. Imprint Academic. [MK]
Kurthen, M., Grunwald, T. & Elger, C. E. (1998) Will there be a neuroscientific theory of consciousness? *Trends in Cognitive Sciences* 2:229–34. [MK]
Lackner, J. R. & Garrett, M. F. (1972) Resolving ambiguity: Effects of biasing context in the unattended ear. *Cognition* 1:359–72. [aGO]
Làdavas, E., Paladini, R. & Cubelli, R. (1993) Implicit associative priming in a patient with left visual neglect. *Neuropsychologia* 31:1307–20. [rGO, MZ]
LeDoux, J. E. (1996) *The emotional brain.* Simon and Schuster. [GW]
Lehky, S. R. & Sejnowski, T. J. (1990) Neural network model of visual cortex for determining surface curvature from images of shaded surfaces. *Proceedings of the Royal Society of London B* 240:51–78. [rGO]
Lestienne, R (1994) Frequency insensitive measures of temporal correlations in spike trains. *Dynamics of Neural Processing International Symposium. Extended Abstract Book* 68–72. [DG]
Levine, J. (1983) Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64:354–61. [aGO]
(1993) On leaving out what it is like. In: *Consciousness: Psychological and philosophical essays,* ed. M. Davies & G. Humphreys. Blackwell. [aGO]
Lewicki, P. (1986) *Nonconscious social information processing.* Academic Press. [rGO]
Lewis, D. (1990) What experience teaches. In: *Mind and cognition,* ed. W. Lycan. Blackwell. [rGO]
Libet, B., Alberts, W. W., Wright, E. W., Delattre, L. D. & Feinstein, B. (1964) Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex. *Journal of Neurophysiology* 27:546–78. [JGT]
Lloyd, D. (1988) Connectionism in the golden age of cognitive science. *Behavioral and Brain Sciences* 11:42–43. [BM, aGO]
(1989) *Simple minds.* Bradford Books/MIT Press. [DL]
(1991) Leaping to conclusions: Connectionism, consciousness, and the computational mind. In: *Connectionism and the philosophy of mind,* ed. T. Horgan & J. Tienson. Kluwer. [DL, BM, arGO]
(1992) Towards an identity theory of consciousness. *Behavioral and Brain Sciences* 15(2):215–16. [DL, EMA]
(1994) Connectionist hysteria: Reducing a Freudian case study to a network model. *Philosophy, Psychiatry, and Psychology* 1(2):69–88. [DL]
(1995a) Consciousness: A connectionist manifesto. *Minds and Machines* 5:161–85. [DL, BM, arGO]
(1995b) Access denied. *Behavioral and Brain Sciences* 18(2):261–62. [DL]
(1996) Consciousness, connectionism, and cognitive neuroscience: A meeting of the minds. *Philosophical Psychology* 9:61–79. [DL, BM, arGO, GNR, MSCT]
(1997) Consciousness and its discontents. *Communication and Cognition* 30(3/4):273–85. [DL]
(1998) The fables of Lucy R.: Association and dissociation in neural networks. In: *Connectionism and psychopathology,* ed. D. Stein. Cambridge University Press. [DL]
Lu, Z. L., Williamson, S. J. & Kaufman, L. (1992) Behavioural lifetime of human auditory sensory memory predicted by physiological measures. *Science* 258:1668–70. [JGT]
MacKay, D. G. (1973) Aspects of a theory of comprehension, memory and attention. *Quarterly Journal of Experimental Psychology* 25:22–40. [aGO]
(1990) Perception, action, and awareness: A three-body problem. In: *Relationships between perception and action,* ed. O. Neumann & W. Prinz. Springer-Verlag. [RAC]
MacLeish, A. (1952) *Collected poems.* Houghton-Mifflin. [DL]
Mandler, G. (1985) *Cognitive psychology: An essay in cognitive science.* Erlbaum. [aGO, GW]
Mangan, B. (1991) Meaning and the structure of consciousness: An essay in psycho-aesthetics. Doctoral dissertation, University of California, Berkeley. [BM]
(1993a) Dennett, consciousness, and the sorrows of functionalism. *Consciousness and Cognition* 2:1–17. [DCD, BM, aGO]
(1993b) Taking phenomenology seriously: The "fringe" and its implications for cognitive research. *Consciousness and Cognition* 2:89–108. [BM, arGO, MSCT]
(1993c) Some philosophical and empirical implications of the fringe. *Consciousness and Cognition* 2:142–54. [BM]
(1996) Against functionalism: Consciousness as an information-bearing medium. Paper presented at the Second Tucson Conference on Consciousness, "Toward a science of consciousness – Tucson II," University of Arizona, Tucson, April 8–13. [DCD, aGO]
(1998) Against functionalism: Consciousness as an information bearing medium. In: *Toward a science of consciousness: The second Tucson discussions and debates,* ed. S. Hameroff, A. Kaszniak & A. Scott. MIT Press. [BM]
Marcel, A. J. (1983) Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology* 15:197–237. [arGO, JP]
(1993) Slippage in the unity of consciousness. In: *Experimental and theoretical studies of consciousness, Ciba Foundation Symposium 174,* ed. G. R. Block & J. Marsh. Wiley. [rGO]
Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information.* Freeman. [rGO, JS]
Massaro, D. (1988) Some criticism of connectionist models of human performance. *Journal of Memory and Language* 27:213–34. [DED]
Masson, M. E. J. (1995) A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21:509–14. [MZ]
Mathis, D. W. & Mozer, M. C. (1995) On the computational utility of consciousness. In: *Advances in neural information processing systems 7,* ed. G. Tesauro, D. S. Touretzky & T. K. Keen. MIT Press. [AC]
(1997) Conscious and unconscious perception: A computational theory. In: *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society.* Erlbaum. [AC]
McClelland, J. L. (1979) On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review* 86:287–330. [AC]
(1995) Constructive memory and memory distortions: A parallel-distributed processing approach. In: *Memory distortion,* ed. D. L. Schacter. Harvard University Press. [DED]
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. [arGO]
McCloskey, M. (1991) Networks and theories: The place of connectionism in cognitive science. *Psychological Science* 6:387–95. [DED]
McRae, K., de Sa, V. R. & Seidenberg, M. S. (1997) On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General* 126:99–130. [MZ]
Meadows, J. C. (1974) Disturbed perception of colours associated with localized cerebral lesions. *Brain* 97:615–32. [RWK]
Merikle, P. M. (1992) Perception without awareness. *American Psychologist* 47:792–95. [GW]
Milner, A. D. & Goodale, M. A. (1993) Visual pathways to perception and action. In: *Progress in brain research, vol. 95,* ed. T. P. Hicks, S. Molotchnikoff & T. Ono. Elsevier. [rGO, JP]
(1995) *The visual brain in action.* Oxford University Press. [RWK, rGO]
Milner, A. & Rugg, M., eds. (1992) *The neuropsychology of consciousness.* Academic Press. [aGO, MZ]

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Mowbray, G. W. (1964) Perception and retention of verbal information presented during auditory shadowing. *Journal of the Acoustical Society of America* 36:1459–64. [aGO]
Murphy, S. T. & Zajonc, R. B. (1993) Affect, cognition and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology* 64:723–39. [GW]
Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83:435–50. [arGO. MV]
Neely, J. H. (1991) Semantic priming effects in visual word recognition: A selective review of current findings and theories. In: *Basic processes in reading,* ed. D. Besner & G. W. Humphreys. Erlbaum. [MZ]
Neisser, U. (1988) Five kinds of self-knowledge. *Philosophical Psychology* 1:35–59. [RAC]
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. [aGO]
Nemirow, L. (1990) Physicalism and the cognitive role of acquaintance. In: *Mind and cognition,* ed. W. Lycan. Blackwell. [rGO]
Neumann, O. & Klotz, W. (1994) Motor responses to nonreportable, masked stimuli: Where is the limit of direct parameter specification? In: *Attention and performance XV: Conscious and nonconscious information processing,* ed. C. Umiltà & M. Moscovitch. MIT Press. [JP]
Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83. [aGO]
Newman, J. (1995) Thalamic contributions to attention and consciousness. *Consciousness and Cognition* 4:172–93. [aGO]
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. [aGO]
Newton, N. (1989) Visualizing is imagining seeing: A reply to White. *Analysis* 49(2):77–81. [NN]
(1996) *Foundations of understanding.* John Benjamins. [NN]
Nolan, K. A. & Caramazza, A. (1982) Unconscious perception of meaning: A failure to replicate. *Bulletin of the Psychonomic Society* 20:23–26. [aGO]
O'Brien, G. (1993) The connectionist vindication of folk psychology. In: *Folk psychology and the philosophy of mind,* ed. S. Christensen & D. Turner. Erlbaum. [aGO]
O'Brien, G. J. (forthcoming) Connectionism, analogicity and mental content. *Acta Analytica.* [rGO]
O'Brien, G. J. & Opie, J. (1997) Cognitive science and phenomenal consciousness: A dilemma, and how to avoid it. *Philosophical Psychology* 10:269–86. [rGO]
(1998) The disunity of consciousness. *The Australasian Journal of Philosophy* 76:378–95. [rGO]
(forthcoming) A defense of Cartesian materialism. *Philosophy and Phenomenological Research.* [rGO]
Opie, J. (1998) *Consciousness: A connectionist perspective.* Ph.D. thesis, University of Adelaide. [rGO]
Palmer, S. (1978) Fundamental aspects of cognitive representation. In: *Cognition and categorization,* ed. E. Rosch & B. Lloyd. Erlbaum. [arGO]
Penrose, R. (1989) *The emperor's new mind.* Penguin Books. [aGO]
Perenin, M. T. (1978) Visual function within the hemianopic field following early cerebral hemidecortication in man. II. Pattern discrimination. *Neuropsychologia* 16:697–708. [aGO]
Perenin, M. T. & Jeannerod, M. (1975) Residual vision in cortically blind hemifields. *Neuropsychologia* 13:1–7. [aGO]
Perkins, D. N. (1981) *The mind's best work.* Harvard University Press. [rGO]
Perruchet, P. & Gallego, J. (1997) A subjective unit formation account of implicit learning. In: *How implicit is implicit learning?,* ed. D. C. Berry. Oxford University Press.
Perruchet, P. & Pacteau, C. (1990) Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General* 119:264–75. [aGO]
Perruchet, P. & Vinter, A. (1997) Learning and development: The implicit knowledge assumption reconsidered. In: *Handbook of implicit learning,* ed. M. A. Stadler & P. A. Frensch. Sage Publications. [AC]
(1998a) Learning and development. In: *Implicit learning: Representation and process,* ed. P. Frensch & M. Stadler. Erlbaum. [rGO, AV]
(1998b) PARSER: A model for word segmentation. *Journal of Memory and Language* 39:246–63. [AV]
Phaf, R. H., Mul, N. M. & Wolters, G. (1994) A connectionist view on dissociations. In: *Attention and performance XV: Conscious and nonconscious information processing,* ed. C. Umiltà & M. Moscovitch. MIT Press. [GW]
Phaf, R. H. & Wolters, G. (1997) A constructivist and connectionist view on conscious and nonconscious processes. *Philosophical Psychology* 10:287–307. [GW]
Plaut, D. C. (1995) Associative and semantic priming in a distributed attractor network. In: *Proceedings of the 17th Annual Conference of the Cognitive Science Society.* Erlbaum. [MZ]
Posner, M. I. & Rothbart, M. K. (1992) Attentional mechanisms and conscious experience. In: *The neuropsychology of consciousness,* ed. A. D. Milner & M. D. Rugg. Academic Press. [RE]
Purcell, D. G., Stewart, A. L. & Stanovich, K. K. (1983) Another look at semantic priming without awareness. *Perception and Psychophysics* 34:65–71. [aGO]
Pylyshyn, Z. W. (1980) Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3:111–69. [aGO]
(1984) *Computation and cognition.* MIT Press. [HC, aGO]
(1989) Computing in cognitive science. In: *Foundations of cognitive science,* ed. M. Posner. MIT Press. [aGO]
Pynte, J., Do, P. & Scampa, P. (1984) Lexical decisions during the reading of sentences containing polysemous words. In: *Preparatory states and processes,* ed. S. Komblum & J. Requin. Erlbaum. [MV]
Ramsey, W., Stich, S. & Rumelhart, D. E., eds. (1991) *Philosophy and connectionist theory.* Erlbaum. [aGO]
Reber, A. S. (1967) Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 5:855–63. [aGO]
(1993) *Implicit learning and tacit knowledge.* Oxford University Press. [rGO]
(1997) How to differentiate implicit and explicit modes of acquisition. In: *Scientific approaches to consciousness,* ed. J. D. Cohen & J. W. Schooler. Erlbaum. [MV]
Reeke, G. N., Jr. (1996) Book review: Patricia S. Churchland and Terrance J. Sejnowski, *The computational brain. Artificial Intelligence* 82:381–91. [GNR]
Rey, G. (1992) Sensational sentences. In: *Consciousness: Psychological and philosophical essays,* ed. M. Davies & G. Humphreys. Blackwell. [aGO]
Robertson, I. H. & Marshall, J. C., eds. (1993) *Unilateral neglect: Clinical and experimental studies.* Erlbaum. [MZ]
Roediger, H. L. & McDermott, K. B. (1993) Implicit memory in normal human subjects. In: *Handbook of neuropsychology, vol. 8,* ed. F. Boller & I. Grafman. Elsevier. [GW]
Rorty, R. (1993) Holism, intrinsicality, and the ambition of transcendence. In: *Dennett and his critics,* ed. B. Dahlbohm. Blackwell. [MK]
Rozeboom, W. W. (1972) Problems in the psycho-philosophy of knowledge. In: *The psychology of knowing,* ed. J. R. Royce & W. W. Rozeboom. Gordon and Breach. [EMA]
Rubel, L. A. (1989) Digital simulation of analog computation and Church's thesis. *Journal of Symbolic Logic* 54:1011–17. [aGO]
Rumelhart, D. E. (1989) The architecture of mind: A connectionist approach. In: *Foundations of cognitive science,* ed. M. Posner. MIT Press. [aGO]
Rumelhart, D. E. & McClelland, J. L., eds. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations.* MIT Press. [aGO]
Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in PDP models. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models,* ed. J. L. McClelland & D. E. Rumelhart. MIT Press. [aGO]
Sacks, O. (1985) *The man who mistook his wife for a hat.* Picador. [RWK, aGO]
(1996) *An anthropologist on Mars.* Picador. [aGO]
Schachter, D. (1989) On the relation between memory and consciousness: Dissociable interactions and conscious experience. In: *Varieties of memory and consciousness: Essays in honor of Endel Tulving,* ed. H. Roediger & F. Craik. Erlbaum. [aGO]
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. [aGO]
Searle, J. R. (1983) *Intentionality.* Cambridge University Press. [aGO]
(1992) *The rediscovery of mind.* MIT Press. [AC, rGO, JO]
Sejnowski, T. J. (1986) Open questions about computation in cerebral cortex. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models,* ed. J. L. McClelland & D. E. Rumelhart. MIT Press. [aGO]
Sejnowski, T. J. & Rosenberg, C. (1987) Parallel networks that learn to pronounce English text. *Complex Systems* 1:145–68. [arGO]
Shallice, T. (1988a) *From neuropsychology to mental structure.* Cambridge University Press. [aGO]
(1988b) Information-processing models of consciousness: Possibilities and problems. In: *Consciousness in contemporary science,* ed. A. Marcel & E. Bisiach. Clarendon Press. [aGO]
(1997) Modularity and consciousness. In: *The nature of consciousness,* ed. N. Block, O. Flanagan & G. Güzeldere. MIT Press. [rGO]
Shanks, D. R. & St. John, M. F. (1994) Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences* 17:367–447. [AC, arGO, MV]
Shepard, R. & Chipman, S. (1970) Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology* 1:1–17. [rGO]

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press

Shepard, R. & Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science* 171:701–03. [rGO]
Smith, S. M. & Blankenship, S. E. (1989) Incubation effects. *Bulletin of the Psychonomic Society* 27:311–14. [rGO]
(1991) Incubation and the persistence of fixation in problem solving. *American Journal of Psychology* 104:61–87. [rGO]
Smolensky, P. (1987) The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy* 26(Supplement):137–61. [aGO]
(1988a) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–23. [arGO]
(1988b) Putting together connectionism - again. *Behavioral and Brain Sciences* 11:59–74. [EMA]
(1991) Tensor product variable binding and the representation of symbolic structures in connectionist systems. In: *Connectionist symbol processing,* ed. G. Hinton. MIT Press. [JS]
Sterelny, K. (1990) *The representational theory of mind.* Blackwell. [aGO]
Stoerig, P. & Cowey, A. (1995) Visual-perception and phenomenal consciousness. *Behavioural Brain Research* 71:147–56. [RWK]
Stoerig, P., Kleinschmidt, A. & Frahm, J. (1998) No visual response in denervated V1: High-resolution functional magnetic resonance imaging of a blindsight patient. *Neuroreport* 9:21–25. [RWK]
Strawson, G. (1994) *Mental reality.* MIT Press. [aGO]
Swinney, D. A. (1979) Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18:645–59. [MV]
(1982) The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In: *Perspectives on mental representation,* ed. J. Mehler, E. C. T. Walker & M. Garrett. Erlbaum. [MV]
Swoyer, C. (1991) Structural representation and surrogative reasoning. *Synthese* 87:449–508. [arGO]
Taylor, J. G. (1996) A competition for consciousness? *NeuroComputing* 11:271–96. [JGT]
(1998a) *The race for consciousness.* MIT Press. (forthcoming) [JGT]
(1998b) Cortical activity and the explanatory gap (with commentaries). *Consciousness and Cognition* 7:107–49; 216–37. [JGT]
Thorndike, E. L. (1949) *Selected writings from a connectionist psychology.* Appleton- Century-Crofts. [DED]
Tienson, J. L. (1987) An introduction to connectionism. *Southern Journal of Philosophy* 26(Supplement):1–16. [aGO]
Tovee, M. J. (1996) *An introduction to the visual system.* Cambridge University Press. [rGO]
Tye, M. (1983) On the possibility of disembodied existence. *Australasian Journal of Philosophy* 61(3):275–82. [NN]
(1992) Visual qualia and visual content. In: *The contents of experience,* ed. T. Crane. Cambridge University Press. [aGO]
(1996) The function of consciousness. *Nous* 30:287–305. [aGO]
(1997) A representational theory of pains and their phenomenal character. In: *Essays on consciousness,* ed. N. Block, O. Flanagan & G. Guveldere. MIT Press. [aGO]
Umiltà, C. (1988) The control operations of consciousness. In: *Consciousness in contemporary science,* ed. A. Marcel & E. Bisiach. Clarendon Press. [aGO]
Umiltà, C. & Zorzi, M. (1995) Consciousness does not seem to be linked to a single neural mechanism. *Behavioral and Brain Sciences* 18:701–02. [MZ]
Van Gelder, T. (1990) Compositionality: A connectionist variation on a classical theme. *Cognitive Science* 14:355–84. [aGO]
(1991) What is the "D" in "PDP"? In: *Philosophy and connectionist theory,* ed. W. Ramsey, S. P. Stich & D. E. Rumelhart. Erlbaum. [HC]
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. [aGO]
Velmans, M. (1991) Is human information processing conscious? *Behavioral and Brain Sciences* 14(4):651–726. [MV]
(1993) Consciousness, causality and complementarity. *Behavioral and Brain Sciences* 16(2):409–16. [MV]
(1996) Consciousness and the "causal paradox." *Behavioral and Brain Sciences* 19(3):537–42. [MV]
(1998) Goodbye to reductionism. In: *Toward a science of consciousness II: The second Tucson discussions and debates,* ed. S. Hameroff, A. Kaszniak & A. Scott. MIT Press. [MV]
Vision, G. (1998) Blindsight and philosophy. *Philosophical Psychology* 11:137–59. [rGO]
Von Eckardt, B. (1993) *What is cognitive science?* MIT Press. [aGO]
Weiskrantz, L. (1980) Varieties of residual experience. *Quarterly Journal of Experimental Psychology* 32:365–86. [aGO]
(1986) *Blindsight: A case-study and implications.* Clarendon Press. [aGO]
(1987) Residual vision in a scotoma: A follow-up study of "form" discrimination. *Brain* 110:77–92. [JP]
Weiskrantz, L., Barbur, J. L. & Sahraie, A. (1995) Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (VI). *Proceedings of the National Academy of Sciences USA* 92:6122–26. [RWK, rGO]
Weiskrantz, L., Warrington, E., Sanders, M. & Marshall, J. (1974) Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain* 97:709–28. [aGO]
Williams, B. (1973) Imagination and the self. In: *Problems of the self.* Cambridge University Press. [NN]
Zeki, S. (1993) *A vision of the brain.* Blackwell. [RWK, arGO]
Zeki, S. & ffytche, D. H. (1998) The Riddoch syndrome: Insights into the neurobiology of conscious vision. *Brain* 121:25–45. [RWK, rGO]
Zihl, J., von Cramon, D. & Mai, N. (1983) Selective disturbance of movement vision after bilateral brain damage. *Brain* 106:313–40. [RWK]

https://doi.org/10.1017/S0140525X99411793 Published online by Cambridge University Press