Consciousness:
A Connectionist Perspective

Jonathan Opie

Department of Philosophy
University of Adelaide

Submitted for the degree of Doctor of Philosophy
February 1998

Contact me at:

Department of Philosophy
University of Adelaide
South Australia 5005

DEDICATION

To my father, who got me thinking, and to Tricia, who provided the love, support, and encouragement that enabled me to see this through.

TABLE OF CONTENTS

ABSTRACT

Cognitive scientists seeking a computational account of consciousness almost universally opt for a *process* theory of some kind: a theory that explains phenomenal experience in terms of the computational processes defined over the brain's representational vehicles. But until recently cognitive science has been dominated by the *classical* computational theory of mind. Today there is a new player on the scene, *connectionism*, which takes its inspiration from a computational framework known as *parallel distributed processing* (PDP). It is therefore appropriate to ask whether connectionism has anything distinctive to say about consciousness, and in particular, whether it might challenge the dominance of process theories.

I argue that connectionism has the resources to hazard a *vehicle* theory of consciousness. A vehicle theory places consciousness right at the focus of cognition by identifying it with the explicit representation of information in the brain. Classicism can't support such a theory because it is committed to the existence of explicit representations whose contents are not phenomenally conscious.

The connectionist vehicle theory of consciousness aligns phenomenal experience with stable patterns of activation in neurally realised PDP networks. It suggests that consciousness is an amalgam of phenomenal elements, both sensory and non-sensory, and the product of a multitude of consciousness-making mechanisms scattered throughout the brain. This somewhat unorthodox picture is supported, I claim, by careful analysis of experience, and by the evidence of the neurosciences.

One obstacle facing this account is the apparent evidence, both direct and indirect, for the activity of unconscious explicit representations in human cognition. I establish that much of the direct evidence for this thesis is open to doubt on methodological grounds. And studies that support the dissociation thesis *in*directly, by way of an inference to the best explanation, are vulnerable to alternative connectionist explanations of the relevant phenomena.

What is most significant about the connectionist vehicle theory of consciousness is not the fact that it's a *connectionist* theory of consciousness, but that it's a *vehicle* theory – an account which takes cognitive science into largely unexplored territory, but in so doing brings into clearer focus the issues with which any theory of consciousness must contend.

STATEMENT OF ORIGINALITY

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

ACKNOWLEDGMENTS

Many thanks to Greg O'Hair, for introducing me to cognitive science, and for much ongoing support, insight and encouragement. Thanks also to the participants in our cognitive science discussion groups – Diarmuid Crowley, Steve Crowley, Greg Currie, Craig Files, Denise Gamble, Gerard O'Brien, Greg O'Hair, Belinda Paterson, Vladimir Popescu and Ian Ravenscroft – for creating a truly stimulating environment in which to hatch ideas, mad and otherwise. Particular thanks to Gerard O'Brien, for providing an inspiring model of how philosophy should be practised, and for unwavering and generous support (I think I owe you a few coffees too!).

Some of the material in Sections 2.2, 2.3, 5.3, 6.1-6.3, 7.1 & 7.2 of this thesis is adapted from:

> O'Brien, G. & Opie, J. (1999) A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* **22**: 127-48

Some of the material in Section 7.4 is adapted from:

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

The material in Chapter 4 forms the basis of:

> O'Brien, G. & Opie, J. (1998) The disunity of consciousness. *The Australasian Journal of Philosophy* **76**: 378-95

1

Introduction

Conscious experience is both wonderful and diverse. It is a subtle blend of elements that not only reflect the condition of our bodies, and the world outside, but also accompany our thought processes, from the most abstract to the most mundane. These phenomenal states are the very stuff of us. While they defy complete description or analysis, they all have something in common – there is *something it is like* to be in them.[1]

The existence of conscious experience is the most pervasive, immediate and compelling reason to believe in the mental. We observe that the world appears a certain way, and recognise this appearing as a special kind of thing, a feature of the natural order that depends on us, and more particularly, on our brains. That is, we recognise the peculiar relationship between the appearance of things and our bodies – appearances inhere in us, even as they reveal an order of objects beyond us. For most of us the mental consists in the stuff of this first-person perspective, even if we acknowledge that there are third-person ways of accessing the mental (by, say, observing behaviour, or measuring brain states).

I will call this the *consciousness criterion* for mentality. It is the view that conscious experience is the fundamental mark of the mental. From this perspective "it may even appear mysterious what sort of thing a mental state might be if it is not a conscious state" (Rosenthal 1986, p.462). The consciousness criterion has long held sway in both popular and critical thinking about the mind. But the second half of this century has seen a highly successful challenge to this view: *intentionality* has come to the fore as the characteristic that sets mental phenomena apart from other kinds of phenomena. According to the *intentionality criterion* it is "aboutness" – the property of standing in the relation of representation to represented – that is the ultimate mark of the mental. Consciousness is thereby rendered a secondary and inessential feature of mental states – a hitchhiker that sometimes goes along for the ride.

The ascendancy of the intentionality criterion is primarily due to the recent successes of cognitive science. Cognitive science is a multi-disciplinary enterprise with the goal, broadly speaking, of developing concrete proposals about the physical realisation of mental processes. More particularly, cognitive science is in the business of explaining intelligence: the capacity to appropriately adapt one's behaviour to novel or changing environmental conditions.[2] Among cognitive psychologists intelligence is often cast in terms of the capacity to solve problems, so another way of expressing the task of cognitive science is to ask: How are naturally occurring systems configured such that they are capable of representing and solving problems? On either formulation, what unites cognitive scientists is the *computational*

---

[1] In using the expression 'conscious experience' my target is neither *self-consciousness* nor what Block (1995) calls *access-consciousness*. It is, rather, *phenomenal consciousness*: the "what it is like" of experience. I will speak variously of 'phenomenal experience', 'phenomenal consciousness', 'conscious experience', or sometimes just plain 'consciousness', but in each case I refer to the same thing.

[2] "Appropriately" is the key word here, and clearly it needs to be cashed out in terms that don't refer to intelligence. One way to proceed is to classify as appropriate those behaviours that contribute to goals it is reasonable to ascribe to the creature or agent in question.

approach to cognition, the attempt to answer this kind of question in terms of disciplined operations defined over in-the-head representations of the world. The idea, roughly speaking, is to suppose that intelligent creatures negotiate the world using *internal models*, and that they can adapt to novel conditions because these models are sufficiently detailed and flexible. Because of its focus on explaining intelligent behaviour, cognitive science has, until recently, largely tended to leave the problem of consciousness to one side. Intentionality, on the other hand, is clearly built in at the ground level. Thus, it is no surprise to find that by far the majority of cognitive scientists regard intentionality as criterial for mentality.

Not only does the intentionality criterion sit comfortably with the *modus operandi* of cognitive science, but cognitive science, in its current incarnation, seems inextricably linked to this approach to mentality. To adopt the alternative, to treat consciousness as the mark of the mental, would appear to be a denial of one of the discipline's central tenets: that cognition involves a great many *unconscious* content-bearing states. There is much to recommend the view that cognitive activity is largely unconscious. The processes of thought are *not* transparent to us. What is six multiplied by seven? The answer (usually) arrives with no conscious effort, but also with no indication at all of the means whereby it was achieved. In recent times this view of thought so dominates, that, contrary to the nineteenth century tradition, in which conscious experience was taken to be the subject matter of psychology, it is *unconscious* processes and states that have become the focus of attention. Lashley is thus willing to assert that "*No activity of mind is ever conscious*" (1956, his italics), and, in a similar vein, Fodor claims that "Practically all psychologically interesting cognitive states are unconscious..." (1983, p.86). It appears to be a *sine qua non* of contemporary cognitive science that human cognition is largely a matter of unconscious operations defined over unconscious representations, and, therefore, that mental properties are first and foremost intentional properties.

The situation would seem to be this: either one must adopt the intentionality criterion, or one must forsake the computational approach to cognition. Searle is a well known advocate of the latter alternative, and has spent much time and energy trying to convince us that, even if a brain can be accurately described as a computer, its intentional properties and capacity for producing intelligent behaviour don't have anything to do with computation (see his 1980 and 1990). By far the majority of theorists favour the first alternative, i.e., they take intentionality to be the fundamental mark of the mental. Consequently, there is a tendency to treat consciousness as something of an add on – a feature of the mind to be tackled once the job of explaining the bulk of cognition has been completed.

However, the discussion so far has largely taken place in the context of a discipline dominated by a single computational framework, that of conventional digital computers, and a corresponding approach to cognition: the *classical* computational theory of mind. Today there is a new player on the scene, *connectionism*, which takes its inspiration from a computational framework known as *parallel distributed processing* (PDP). It may be that the dominance of the intentionality criterion is imposed not by the computational theory of mind *per se*, but by the classical formulation of this generic doctrine. How does the situation change when we look at these issues from the connectionist perspective?

It changes dramatically, I will argue. Connectionists are in a position to hazard an account of phenomenal experience that returns consciousness to centre stage. But, significantly, this does not amount to reinstating the consciousness criterion of mentality. Connectionism in fact permits a reconciliation between the two criteria of mentality. It does so by exploiting an obvious and natural relationship between consciousness and

representation (namely, that experience is representational).[3] And it does so without abandoning the computational approach to cognition.

My project, then, is to develop and defend a connectionist theory of phenomenal experience with these features. In particular, I will defend a theory that places consciousness right at the focus of cognition, by identifying it with the explicit[4] representation of information in the brain. I begin (in the next chapter) with a discussion of classicism and connectionism, with the aim of teasing out what is distinctive about these two accounts of cognition. To this end, the discussion will be set within the context of a taxonomy of mental representation developed by Dennett (1982). The three chapters that follow turn to general issues concerning the nature and explanation of phenomenal experience.

Chapter 3 consists of a relatively theory neutral sketch of phenomenal experience, and of the relationship between consciousness and thought. This sketch is intended as a preliminary benchmark against which any proposed theory of consciousness may be measured. I argue, contrary to many treatments, that phenomenal experience is highly *parallel*, both inter-modally and intra-modally. It is a multi-modal aggregate – a composition of relatively independent phenomenal elements – and includes a great deal of what I call *abstract phenomenology*, particularly in the service of higher thought processes.

Chapter 4 addresses the so called "unity of consciousness" and tries to untangle various different senses in which conscious experience might be said to be unified. I first consider (and reject) the view that consciousness is a unity because it comprises a single informational content at each instant. This I call the monophonic model of consciousness. The alternative polyphonic model – which recognises the multi-modal nature of instantaneous consciousness – is far more plausible, but leaves us in need of some account of the unity of consciousness. One option is to assume that the brain realises a single consciousness-making system or mechanism of some kind. I reject this single-track approach on both evolutionary and neuroscientific grounds, and propose instead a multi-track approach to consciousness, namely: that the brain incorporates a multiplicity of consciousness-making sites, such that each identifiable stream of experience has independent origins in the brain. This option becomes available when one recognises that the unity of consciousness doesn't have to be conceived literally in terms of oneness. Phenomenal experience is "unified" in the sense that it is typically coherent, and incorporates a sense of self, or a point of view.

The widely held view, remarked on above, that processes defined over unconscious representations figure in human cognition, creates an obvious difficulty for an account that seeks to identify phenomenal experience with the explicit representation of information in the brain. Indeed, there is a large body of research which appears to favour the view that mental representation and conscious experience are *dissociable*, i.e., that the one can occur in the absence of the other. More particularly, there is evidence to support the stronger claim –

---

[3] It may be that we infer to the existence of internal representations as the best explanation of intelligent behaviour, but what compels acceptance here is not the power of abduction, it is the first-hand *experience* of representations. The world is represented *through* the medium of consciousness, and the appearance of an object is the phenomenal aspect of its being represented.

[4] It is commonplace for theorists to distinguish between *explicit* and *inexplicit* 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 inexplicitly. See, e.g., Dennett 1982; Pylyshyn 1984; and Cummins 1986. I discuss the distinction between explicit and inexplicit representation more fully in Chapter 2.

which I call the *dissociation thesis* – that *explicitly represented* unconscious information has a role in human cognition.

In Chapter 5 I look at the origins of, and support for, the dissociation thesis, before considering some doubts that have been raised about it. I begin by showing how the dissociation thesis relates to the so called "limited capacity" of consciousness (much discussed by cognitive psychologists), and then explore the reasons, both historical and empirical, for its general acceptance. Empirical studies supportive of the dissociation thesis divide into two broad classes: those which provide good *direct* evidence for the presence of unconscious explicit representations in cognitive processes; and those which, because couched in its terms, *indirectly* support the dissociation thesis (by way of an inference to the best explanation). I will suggest that experimental analyses which take the dissociation thesis for granted are typically open to independently plausible explanations that eschew unconscious, explicit representations, and that many of the studies which putatively provide good direct evidence for the thesis suffer from methodological flaws of one kind or another. My aim here is not to develop a thorough-going refutation of the dissociation thesis, but simply to undermine the widespread view that the thesis is obligatory for cognitive scientists.

Having laid some groundwork concerning the nature and explanation of consciousness, I spend the remainder of the thesis examining the prospects for a connectionist theory of consciousness.

My aim in Chapter 6 is to demonstrate how issues of fine-grained computational architecture impact on the explanation of consciousness. I begin by distinguishing two distinctive kinds of theory that are available to computational theorists when it comes to explaining consciousness:

- *vehicle theories*, which seek to identify conscious experiences with the vehicles of explicit representation in the brain; and
- *process theories*, which take consciousness to emerge from the computational activities in which (some of) those vehicles engage.

Vehicle theories are incompatible with the dissociation thesis, while process theories implicitly assume it to be sound. I go on to demonstrate that there is a close (and principled) link between the dissociation thesis and the classical computational theory of mind, and thus that classicists have no choice but to adopt a process theory of consciousness. What has not been widely appreciated is that connectionism, because it relies on the distinctive representational and computational resources of parallel distributed processing (PDP) systems, is able to dispense with the dissociation thesis. Consequently, connectionists are in a position to hazard a vehicle theory of phenomenal experience.

Establishing this possibility for connectionism doesn't amount to demonstrating that such a theory is satisfactory. However, since vehicle theories are all but absent in contemporary cognitive science, it is vital that this much neglected region of the theoretical landscape be opened up for serious exploration. I take up this task in Chapter 7, where I develop and defend the conjecture that consciousness is identical to the explicit representation of information in the brain, in the form of stable patterns of activation across neurally-realised PDP networks. I show that this hypothesis can be used to account for the various features of phenomenal experience identified in Chapter 3. It has the capacity to explain the great diversity in experience by subsuming both *intra*modal and *inter*modal experiential differences under a single activation space account, and it can account for the various degrees of abstractness exhibited by the elements of experience in terms of the hierarchical organisation of the brain. The connectionist vehicle theory of consciousness is

best interpreted, I argue, as a multi-track theory of consciousness, and so can also do justice to the various kinds of evidence supportive of that approach presented in Chapter 4.

A special problem for a vehicle theory of consciousness is the need to take up the explanatory burden carried by explicit representations in conventional accounts of unconscious processing. I devote considerable space to this issue, and argue that connectionism does have the resources to make good on a denial of the dissociation thesis.

I conclude the chapter by considering the relationship between phenomenal consciousness and informational access from the perspective of the present account.

What is most significant about the connectionist vehicle theory of phenomenal experience is not the fact that it's a connectionist theory of consciousness, but that it's a vehicle theory – an approach to consciousness made possible by the unique representational and computational features of the PDP framework. This approach takes cognitive science into new territory, and opens up the prospect of a computational theory which returns consciousness to its former place at the centre of mentality. But in so doing the connectionist vehicle theory does not lead to the consciousness criterion of mentality. Rather, it offers a way of bringing mental representation and phenomenal experience together without privileging either aspect of mentality over the other, and without abandoning the computational theory of mind.

2

Two Computational Theories of Mind

Two fundamentally different accounts of mental architecture and the nature of mental processes are currently vying for the computational high ground in cognitive science. Classicism established a strong position early in the piece, but connectionists have lately regrouped, and the combatants are now too well matched to confidently predict how things will turn out. In this chapter I'll introduce these two computational theories of mind. My discussion will include a brief examination of the characteristic strengths and weaknesses of both classicism and connectionism. In order to provide a framework within which their distinctive features may be best exhibited, the discussion will be set within the context of a taxonomy of mental representation developed by Dennett (1982). This taxonomy will provide a grounding for much of the material in subsequent chapters. To begin, I turn to the conceptual origins of the computational approach to cognition.

2.1 Folk Psychology and Cognition

For better or worse cognitive science has developed under the more or less overt influence of the philosophy of mind, and the approach to cognition presupposed by theorists working in that tradition. Philosophers have characteristically tended to focus on the implicit theory of thought embodied in folk explanations of human behaviour[5]. And, according to many philosophers, what the folk are telling us is that thoughts are *propositional attitudes*: mental states that can be canonically described in sentences with an intentional verb (such as believes, hopes, remembers) followed by a that-clause containing a proposition. For example:

She wanted to go to the ball.

She hoped that he would come and collect her.

She remembered that his carriage was damaged.

These can be either *occurrent* or *dispositional*, meaning that they can properly be ascribed not only to a thinker who is currently entertaining, or affected by, the thought in question (it is then said to be occurrent), but also to a thinker for whom the thought is currently inactive. Thus, it is perfectly reasonable to say of a person in deep sleep: "He believes that the carriage will be repaired". In so doing one is ascribing a dispositional thought.

Thought is being treated here in essentially static terms, as a relation between subject and proposition. But the term 'thought' also admits of a dynamic reading. We not only say: "What are your thoughts on the matter?", referring to the results of thinking, but also: "She is deep in thought", referring to the process of generating thoughts. In other words, the folk

[5] More specifically, the subclass of human behaviours that can best be made sense of in terms of the activities of an agent who guides behaviour so as to satisfy current goals in light of current conditions. This looks a bit circular, but I think it actually excludes quite a few human behaviours (such as simple reflex responses), since these can *best* be understood in terms of a neurological/anatomical story – folk psychology adds nothing useful here. The principle seems to be this: invoke the explanatory framework of folk psychology when lower-order explanations are not forthcoming.

also have a story about thinking – about the generation of thoughts (conceived as propositional attitudes), and the causal relations between thoughts, perceptions and actions. Consider the following:

> When she saw the clock she realised it was nearly time for the ball. Since she desperately wanted to go she hurriedly started to dress. Suddenly she remembered that his carriage was damaged, and feared he would not come to collect her.

Seeing the clock generates a belief (that the time is so-and-so). Wanting to go to the ball, in conjunction with this and various other beliefs, and the desire to look nice, leads to the activity of getting dressed. Implicit here is a causal economy, in which perceptions generate thoughts, thoughts interact to produce further thoughts, and some of these issue in appropriate actions. Moreover, thinking, on this account, is *rational*; fearing he will not come *makes sense* in light of the belief that his carriage is damaged (given reasonable background assumptions).

Common parlance thus presupposes a rich framework with which to explain and predict human behaviour. The framework in question is belief-desire (or propositional attitude) psychology, roughly the view that:

- belief fixation begins with one or another form of perception, and generates new beliefs from old in ways that respect the semantic relations between their respective contents;
- human action issues from the interaction of beliefs and desires, and typically seeks to realise the latter, in light of the former.

While there is plenty here in need of development and clarification, the central idea is this: human behaviour is largely the result of sense-preserving interactions among a set of mental representations.

Clearly, this characterisation of folk psychology is biased towards a causal interpretation – an interpretation that treats beliefs, desires, and so forth as internal, behaviour-causing states. Such an interpretation is not to everyone's taste. Philosophical behaviourists one stripe or another, for example, have tended to treat mental states in a non-causal fashion. Among the behaviourists the vocabulary of folk psychology is understood to refer to behaviours, or dispositions to behave, or instruments in an interpretative theoretical framework, *not* to causally efficacious internal states. That is, mentalistic terms are tolerated only insofar as they are understood to have behavioural criteria, or to get their meaning within an instrumental theory of belief/desire attribution. The *criterial* approach is evident in Wittgenstein's *Philosophical Investigations* (1958), and was championed by Ryle (1949). It is subject to some well known difficulties regarding the characterisation of beliefs (and other mental states) as particular individual dispositions to outward behaviour (see Bechtel 1988, pp.91-3 for discussion). The general consensus today is that individual beliefs and desires are not attributable in isolation.

More recently, Davidson (via Quine) has developed an approach in which the business of attributing mental states is conceived as a thoroughly holistic theoretical enterprise; it is a matter of putting in place a *system* that renders coherent an individual's behaviour (see, for example, his 1974). This is carried out according to principles of charity and rationality, and is modelled on the approach one must take in translating a radically unfamiliar language. The set of beliefs and desires so generated forms a rational scheme which we can use to explain and predict behaviour. Davidson's approach to the mental talk is thus *interpretationist*, because he believes there is nothing more to possessing any given belief or desire than being interpretable in the light of an optimal scheme of thoughts containing it. Dennett (1987, 1991a) defends a similar view. According to Dennett the apparatus of

belief/desire psychology provides, not a causal theory, but a schematisation of behaviour. Using this apparatus amounts to adopting a particular explanatory stance: the *intentional stance*, which involves ascribing beliefs and desires according to a *rationality assumption* (roughly that an intentional system will mainly believe true things, and mainly desire what is good for it). From this viewpoint, all there is to being a believer is being *reliably predictable* from the intentional stance. It is *not* a matter of being a container for causally efficacious representational vehicles bearing propositional contents.[6]

It is not my purpose here to argue the relative merits of causal and non-causal interpretations of folk psychology. However, when treated as a causal theory, folk psychology does, I think, contain an important lesson for cognitive science. Recall that cognitive science is in the business of explaining intelligence. It seeks to discover the causal processes that underlie our capacity, and that of many organisms, to respond appropriately to novel or changing environmental conditions (see Chapter 1). Now consider what the folk have to say about cognition. Thinking, they tell us, involves the causal interplay of beliefs, desires and other propositional attitudes. Objects and events in the environment are partly causally responsible for the creation of such states, moreover, these same states are causally implicated in the production of our behaviour – thus, beliefs, desires and so on, are the causal nexus that links environmental conditions to behavioural responses. Within this framework propositional attitudes have a very special status, because they are not only conceived as *in-the-head physical* states, but also as *intentional* states – they are *representations* of external states of affairs. Moreover, in playing the role of mediators between input and output, propositional attitudes interact in such a way that the immediate products of their interactions (further PAs) respect the semantic relations obtaining between the reactants.[7] That is, the causal relations among the representations posited by folk psychology are *semantically coherent*.[8] So the kind of causal story required to explain intelligent behaviour, according to the folk, is one that posits semantically coherent causal operations defined over in-the-head representations. In other words, when we treat folk psychology as a causal theory, what we get is a generic *computational* account of cognition.[9]

What folk psychology teaches us, then, is that the causal processes responsible for intelligent behaviour may be fruitfully construed as computational processes (generically

---

[6] Dennett differs from others in this tradition in his insistence that, while it is a mistake to treat beliefs and desires as in-the-head representations, a mature science of the mind will nevertheless trade in information-bearing states of some kind. Thus, despite his instrumentalism (or "mild realism" (1991a)) "about the belief-states that appear as *abstracta* when one attempts to interpret…real phemonena by adopting the intentional stance", Dennett avers that he is "as staunch a realist as anyone about those core information-storing elements in the brain, whatever they turn out to be, to which our intentional interpretations are anchored" (1987, pp.70-2). This places Dennett squarely in the *cognitive camp*; a position occupied by all those who treat internal representations of some grain as essential to explanations of intelligent behaviour.

[7] Recall the fearing that *he would not come* which was (partly) caused by the remembering that *his carriage was damaged*. See above.

[8] Fodor refers to this as the "parallelism between causal powers and contents" (1987, p.13).

[9] This account is computational, I suggest, because the class of devices that realise semantically coherent causal processes are precisely those we are inclined to interpret in computational terms. To properly support this claim it would be necessary to undertake an extensive survey of devices that are currently classified as computers (including both analog and digital machines) and demonstrate that these are all captured by the semantic coherence account of computation. It would also be necessary to show that such an account excludes a whole host of complex dynamical systems whose behaviour we *don't* typically construe in computational terms. I don't propose to undertake either task here, but simply note that the semantic coherence account is implicit in the writings of a number of prominent theorists (notably Fodor 1975, Ch.1, Pylyshyn 1984, Ch.3, and Haugeland 1985, Ch.3), and has been explicitly defended by Smith (1982) and O'Brien (1993).

conceived).[10] This may come as a surprise to some, given that the dominant origin myth for cognitive science takes, as its starting point, the construction (in the 1940s) of the first electronic digital computers, and then explains how theorists – impressed with this new technology – proceeded to develop a new science of the mind under the guidance of the "computer metaphor". There is something right about this analysis of history, despite the fact that Hobbes explicitly articulated a (digital) computational view of cognition in the 1650s (see Haugeland 1985, pp.23-8 for discussion), and that folk psychology has arguably been trading in implicit computational theory for many thousands of years. Electronic digital computers not only constitute a superb tool for modelling our thought-processes, but they have undoubtedly had a profound effect on the imaginations of theorists in this century, who could, for the first time, see working demonstrations of mechanical reasoning. Nevertheless, it is wrong to suggest that the *concept* of computation is the offspring of modern computer theory and practice – the conceptual lineage is, rather, the reverse. It is awareness that our (conscious) mental processes involve semantically coherent operations over inner representations that has provided the impetus to develop the abstract theory of computation, and to construct both digital and analog computers.[11] As O'Brien puts it:

> These artefacts are the culmination of a conscious attempt to design and build physical systems that instantiate the kinds of causal sequences constituting our mental processes. The devices we construct and call 'computers'…are made in our own image. (1993, p.38)

Thus, to construe minds in computational terms is not to employ a metaphor. As Von Eckardt points out, a metaphor, even when it is "theory-constitutive", remains a nonliteral attribution. To speak of a "computer metaphor" is to suggest that cognitive scientists don't regard computational devices as the kind of thing the mind could literally be. (1993, pp.98-

---

[10] Some feel that it has more to teach us than this. A further moral one could draw from folk psychology concerns the *kinds of contents* that a mature theory of mind should trade in. The concepts and propositions of folk psychology – the "folk solids", as Clark (1993, Ch.1) calls them – are pitched at the level of everyday objects, properties and relations. Friends of the folk argue that the obvious success of folk psychology in accounting for human behaviour warrants a realistic treatment of the folk solids. (Fodor (1978, 1987, Ch.1) is the best known proponent of this view, but see also Dretske 1988, and Horgan & Woodward 1985.) Such theorists expect that cognitive science will eventually deliver a science of the mind that quantifies over commonsense contents. The moral is this: folk psychology is a reliable guide to the sorts of contents that are tokened by (at least some of) the representations implicated in human cognitive activity.

This is the kind of claim made by advocates of the Representational Theory of Mind (RTM). Of course, the idea is not that every content we attribute in ordinary mentalistic discourse must have an inner analogue, nor that every mental representation must correspond to a propositional attitude. Rather, the claim is that among the mental representations countenanced by a mature science of the mind, there will be some whose contents are recognisable by the folk. As Fodor puts it:

> the vindication of belief/desire explanation by RTM does not require that every case common sense counts as the tokening of an attitude should correspond to the tokening of a mental representation, or vice versa. All that's required is that such correspondences should obtain in what the vindicating theory itself takes to be the core cases. (1987, p.24)

RTM also makes an important claim about the attitudes, namely: that they are realised by computational relations between an organism and its mental representations (see, e.g., Fodor 1987, p.17). One could, in principle, accept that some of our mental representations carry folk-level contents, *without* adopting this story about the attitudes. For example, one might favour a *monadic* account, which involves treating attitudes as part of the *contents* of mental representations. Thus, in addition to the issue about mental contents, there is a further issue as to whether the propositional attitudes, construed as relations, pick out a salient feature of the inner economy. Strictly speaking, it is only by accepting the standard account of both that one arrives at RTM.

[11] The Turing machine, for example, was clearly the result of Turing's attempt to demonstrate that a deterministic machine could realise the kinds of (semantically coherent) mental processes he discovered in consciousness when performing arithmetic (Turing 1937).

104) But that can't be right, since cognitive science starts from the recognition, courtesy of the folk, that *we* are computational systems of some kind.

It is the explicit focus on computation that distinguishes cognitive science from all the various disciplines that attempt to come to grips with the mind in some way; a focus that originates with the causal construal of folk psychology. That is, what makes cognitive science special is the *generic computational theory of mind*: the claim that human cognitive processes consist of semantically coherent causal operations defined over contentful, in-the-head states.[12] Armed with this theory, cognitive science is in a position to begin the task of generating concrete models of cognition. However, while the generic computational theory of mind is an intuitively appealing framework within which to develop explanations of intelligence, the kinds of theoretical constraints it puts in place are quite liberal. In particular, this theory still leaves us with a profound and difficult question, namely: "What sort of mechanism could have states that are both semantically and causally connected, and such that the causal connections respect the semantic ones?" (Fodor 1987, p.14). It is now generally accepted that cognitive science countenances two distinct answers to this question: one arising from the theory of digital computers, and another from recent work with parallel distributed processing (PDP) systems. Each of these answers issues in a distinct approach to cognition – the *classical* computational theory of mind, and the *connectionist* computational theory of mind, respectively. It is my task, in what follows, to sketch out these two theories.

2.2 The Classical Computational Theory of Mind

The *classical* program in cognitive science has its origins in the computational theory that underpins the operation of conventional (digital) computers.[13] Classicism's fundamental commitment is to the claim that human cognitive processes are *digital* computational processes. In order to understand classicism it will thus be necessary to briefly examine the theory of digital computers. Once I've provided an explicit formulation of the classical computational theory of mind, I will go on to look at its implications for the nature and kinds of information coding in the brain.

A digital computer is, in Haugeland's well known terms: an *interpreted automatic formal system* (1981, 1985). An automatic formal system is a device, some of whose states correspond to the tokens in a formal system, and whose state transitions are so engineered that they correspond to token manipulations permitted by the rules of the system. To get a physical device to behave in this way is no mean feat, but, roughly speaking, what's required is that it be made sensitive to certain high-level (*syntactic*) properties of the states that realise tokens, such that its operations unambiguously correspond to transformations describable at that level. And the way one pulls *that* off is to arrange things so that minor variations in the physical realisation of individual tokens do not affect the course of processing. Only variations that are discernible at the syntactic level can be allowed to have any predictive value with regard to the long-term behaviour of the system.

---

[12] This way of putting things is due to O'Brien 1993, p.41.

[13] Prominent contemporary philosophers and cognitive scientists who advocate a classical conception of cognition include Chomsky (1980), Field (1978), Fodor (1975, 1987), Harman (1973), Newell (1980), Pylyshyn (1980, 1984), and Sterelny (1990).

An automatic formal system qualifies as a digital computer when it satisfies an *interpretation*, that is, when:

1) there is a mapping between the distinct token types of the system, and the objects, properties and relations of some represented domain, such that;

2) the sequence of token manipulations performed by the system is semantically well-behaved.

The latter condition amounts to a requirement that the token manipulations of the system make reasonable sense, given what the tokens refer to. The beauty of all this is that we already know how to formalise many important domains, such as arithmetic, and argument (to some extent). This means that semantically coherent processes (such as inference) are amenable to treatment as rule-governed operations defined over formal tokens, which operations preserve all the appropriate semantic relations among the represented entities.[14] Since we also know, from computer science, how to automate formal systems (as sketched above), we have at our disposal at least one means of realising semantically coherent causal processes: as *syntactically-governed symbol manipulations*. 'Symbol' here refers to the tokens of an interpreted, automatic formal system. These are entities that can be individuated both semantically (according to their contents), and syntactically (according to certain of their high-level physical properties).

The full generality and power of digital computation only becomes apparent when we allow for the possibility of *complex* symbols. A complex symbol is, by definition, a symbol that is composed of two or more *atomic* symbols (symbols which have no semantically significant parts). Complex symbols bring an added layer of complexity to an automated formal system, both because they require us to specify *composition rules* according to which atomic tokens may be legally combined, and because they rely for their physical realisation on a *combinatorial syntax* – a syntax that is some compositional function of the syntax of simple tokens, together with the manner in which they are to be combined. Moreover, the manipulation rules of the system, together with their physical realisations, must clearly be sensitive to the internal structure of these symbols. Interpreting such a system requires not only that the allowed state transitions be semantically coherent, but that the composition rules of the system produce tokens that make reasonable sense, given the meanings of the simple tokens of which they are composed. What all this buys you is a far more powerful machine, because, with a set of recursive composition rules, it is possible to generate an essentially open-ended set of symbol structures, even with a finite stock of primitive symbols. In other words, a digital computer that employs complex symbols is able to achieve *productivity*.

The *classical computational theory of mind* is the claim that human cognitive processes are digital computational processes. Given what we know about digital computers, this implies that human cognition consists in the manipulation of *mental* symbols – the neurally realised tokens of a (yet to be determined) formal system, or systems. Such symbols will (in all likelihood) be structured entities, with a combinatorial syntax and semantics. In other words, they will realise a *Language Of Thought* (LOT): a set of syntactically defined atomic token types, which combine to form complex token types according to specifiable composition rules, and a systematic semantic mapping between these token types and external objects, properties, relations, and so forth. Putting everything together, classicism may be succinctly characterised as follows.

---

[14] For example, entailment relations among certain classes of propositions are preserved by relations of formal derivability among the (complex) tokens that represent them in predicate logic. See Mates 1965, pp.164-5.

1) There is a language of thought physically realised in our brains.

2) Human cognitive processes are the *syntactically-governed manipulations of symbols* written in this language of thought.[15]

The classicist thus takes the generic computational theory of mind – the claim that human cognition consists of semantically coherent causal processes defined over in-the-head representations – and adds a more precise account of the inner vehicles (they are complex symbols in the language of thought), and the computational processes involved (they are syntactically-governed symbol manipulations).

There are some immediate payoffs for classicism in light of all this. As I've indicated, a digital computer whose state transitions are defined over complex symbols is a potentially productive system. Classicism, because it posits a language of thought, gets a nice explanation of the *productivity of thought* for free. Thought is said to be productive because of its open-endedness: the fact (according to some theorists) that we are, in principle, capable of producing and comprehending an essentially unbounded set of distinct thoughts. For classicists the productivity of thought is a simple consequence of the fact that our occurrent thoughts are tokens in a system of mental representations with a combinatorial syntax and semantics. Such a system, assuming it relies on a set of physically realised recursive rules of combination, is limited only by such factors as memory capacity. (Fodor 1987, Appendix; Fodor & Pylyshyn 1988).

The LOT hypothesis also provides a very straightforward explanation for the *systematicity* of thought. What makes thought systematic is the (alleged) fact that our capacity to think one set of thoughts is *intrinsically related* to our capacity to think other, structurally related thoughts. For example, if one can think Jon loves movies, and Ry hates music, one can surely also think Ry loves movies, and Jon hates music. On the assumption that such thoughts are expressed via an inner symbol system, with a combinatorial syntax and semantics, this fact about or mental life is no mystery. Thoughts like Jon loves movies, and Ry hates music, are not simples; they are composed of parts, and these parts are *recombinable*. Thus, to be able to think the first of the above pairs of thoughts, is to have the wherewithal to think other thoughts that are composed of the same parts (e.g., the second pair).

Classicism and Mental Representation

My next task is to discover what the classical computational theory of mind implies about the forms of information coding in the brain. I will introduce a taxonomy due to Dennett (1982), consisting of four distinct styles of representation, which I believe respects the implicit commitments of most theorists in this area (see Cummins 1986, and Pylyshyn 1984). This taxonomy will prove a useful framework for the comparison and further explication of the classical and connectionist accounts of cognition.

---

[15] The classical source of this characterisation is Fodor 1975 (but see also his 1987, Ch.1). O'Brien (1993) provides a very clear recent explication of this position. One qualification is in order. Given the current popularity of modular theories of mind, where human cognition is taken to be a co-operative activity of a number of semi-independent cognitive agents or "modules" (see, e.g., Fodor 1983, Minsky 1985), it is probably better to characterise classicism as the claim that there are a number of distinct languages of thought physically realized in our brains.

First, Dennett tells us, information can be represented in an *explicit* form:

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

To take a familiar example: in a Turing machine the symbols written on the machine's tape constitute the "physically structured" vehicles of explicitly represented information. These symbols are typically subject to an interpretation (provided by the user of the machine), and can be "read" by virtue of mechanisms resident in the machine's read/write head. They are thus "explicit representations", according to Dennett's taxonomy – physically distinct objects, each carrying a single semantic value.

In the classical context, explicit representation consists in the tokening of symbols in some neurally realised 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. What's more, it is these objects that are responsible for the course of mental processing. As Fodor and Pylyshyn put it:

> 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. (1988, pp.13-14)

It is these "physical counterparts of the symbols" upon which explicit information[16] supervenes, according to the classicist.

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

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

It is questionable, however, whether the concept of implicit representation, defined in this way, is relevant to classical cognitive science. Logical consequences don't have effects unless there are mechanisms whereby a system can *derive* (and *use*) them. And it is clear from the way Dennett defines it that implicit information can exist in the absence of such mechanisms. Another way of putting this is to say that while the information that a system implicitly represents does partly supervene on the system's physical substrate (the explicit tokens that act as premises), its supervenience base also includes principles of inference

[16] In what follows, whenever I 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.

which need not be physically instantiated. Thus implicit representation is really just a logical notion, and not one that can earn its keep in cognitive science.

However, an implication that a system is *capable of drawing*, is a different matter. Dennett refers to information that is not currently explicit, but which a computational system is capable of rendering explicit, as *potentially explicit* (1982, pp.216-217). Representation of this form is not to be unpacked in terms of mere logical entailment, but in terms of a system's computational capacities. For example, a Turing machine is typically capable of rendering explicit a good deal of information beyond that written on its tape. Such additional information, while not yet explicit, isn't merely implicit; it is potentially explicit. And it is potentially explicit in virtue of the symbols written on the machine's tape and the mechanisms resident in its read/write head.[17]

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

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

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

The Turing machine can again be used to illustrate the point. The causal operation of a Turing machine, remember, is entirely determined by the tokens written on the machine's tape together with the configuration of the machine's read/write head. One of the marvellous features of a Turing machine is that computational manipulation rules can be explicitly written down on the machine's tape; this of course is the basis of stored program

---

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

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

In a similar vein, tacit representation is implicated in our primitive cognitive processes, according to the classicist. These operate at the level of the symbolic atoms and are responsible for the transformations among them. No further computational story need be invoked below this level; such processes are just brute physical mechanisms. Classicists conceive them as the work of millions of years of evolution, embodying a wealth of information that has been "transferred" into the genome. They emerge in the normal course of development, and are not subject to environmental influences, except in so far as some aspects of brain maturation require the presence of environmental "triggers". So classical cognition bottoms out at symbolic atoms, implicating explicit information, and the "hardwired" primitive operations defined over them that implicate tacit information. In the classical context we can thus distinguish tacit representation from both explicit and potentially explicit styles of mental representation as follows: of the physical structures in the brain, explicit information supervenes only on tokened symbolic expressions; potentially explicit information supervenes on these structures too, but *also* on the physical mechanisms capable of rendering it explicit; in contrast to both, tacit information supervenes *only* on the brain's processing mechanisms.[18]

2.3 The Connectionist Computational Theory of Mind

In this section I will introduce the connectionist computational theory of mind. 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).[19] In order to bring out what is special about connectionism it will be necessary to examine this framework in some detail. One of my aims here is to show how Dennett's taxonomy of classical representational styles can be usefully mapped onto connectionism, and in a way that illuminates the distinction between classicism and connectionism.

PDP directly models some high-level physical properties of the brain. A PDP network consists in a collection of simple processing units ("neurons"), each of which has a continuously variable activation level (its "spiking frequency"). The complex interaction among real neurons is modelled by connecting processing units with connection lines ("axons"), which enable the activation level of one unit to contribute to the input and subsequent activation level 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

---

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

[19] The *locus classicus* of PDP is the two volume set by Rumelhart, McClelland, and the PDP Research Group (Rumelhart & McClelland 1986, McClelland & Rumelhart 1986).

new activation level that is some threshold function of its present activation level and that sum.

PDP networks compute in one of two ways, depending on their architecture. *Feedforward* networks, consisting of one or more layers of units in which the flow of activation is strictly one-directional, compute by transforming patterns of activation over their input lines into patterns of activation over their output units. Such networks can be employed in pattern association, pattern completion (auto-association), and pattern recognition tasks (Rumelhart, Hinton & McClelland 1986). *Recurrent* or *interactive* networks, which have feedback connections among some of their units (and in the extreme case have fully reciprocal connections), compute by "relaxing" into a stable pattern of activation in response to a stable array of inputs over their input lines. Such networks have the capacity to keep track of long-range dependencies amongst the members of an input sequence (Elman 1989, 1990), model oscillatory phenomena (Jordan 1989), and resolve ambiguities (which amounts to finding the best-fit solution to a problem in which numerous constraints must be simultaneously satisfied).[20] While the achievement of a stable activation level in a feedforward network requires only a few processing cycles, relaxation search can be relatively time consuming, given that it relies on a kind of conflict resolution procedure. In both cases, however, the course of processing is conditioned by the same two factors: (1) the magnitudes of the weights on the connection lines; and (2) the network connectivity structure. These jointly determine the way that activation is passed from unit to unit, and thereby shape the networks' response to the input it receives.

The representational capacities of PDP systems rely on the plasticity of the connection weights between their constituent processing units. By altering the connection weights one alters the activation patterns the network produces in response to its inputs. The magnitudes of these weights are set using some kind of learning rule[21], which rule typically relies on an assessment of the networks' performance relative to the output it *ought* to be producing – that is, relative to the required set of vector transformations. When learning depends on a teaching signal generated outside the PDP system it is said to be *supervised*, and is clearly biologically unrealistic. However, as Churchland and Sejnowski point out (1992, pp.97-8), no such criticism applies to learning procedures that rely on *internal* feedback signals (internal *monitoring* of performance), or on some kind of competitive algorithm.

Through repeated applications of a learning rule, an individual network can be taught to generate a range of stable target patterns in response to a range of inputs. These *activation pattern representations*, as I will call them, are a transient form of information coding in PDP networks, since they constitute a highly volatile response to current input conditions. Such patterns are sometimes treated in a *localist* fashion, in that the activation levels of individual units are used to represent distinct items in the represented domain. However, there is good reason to believe that the brain doesn't generally encode information in this fashion. Neurons appear to code *coarsely*, meaning that individual neurons exhibit sensitivity to a broad range of stimulus conditions (which is not to deny that they are often maximally sensitive to a single stimulus-type). This finding is most consistent with the view that the brain represents information in a *distributed* fashion, i.e., that the *pattern* of stable activation is the basic semantic unit in real nervous systems.[22]

---

[20] See Churchland & Sejnowski 1992, pp.115-25, or Churchland 1995, Ch.5 for more detailed discussion of the properties of recurrent networks.

[21] The most common of these is the *back propagation* algorithm – see, for example, the discussion in Rumelhart, Hinton & Williams 1986.

[22] Churchland & Sejnowski 1989, pp.38-41 presents some fascinating evidence for this claim.

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 being *explicitly* represented. For these patterns are causally efficacious, physically structured events, embedded in a system with the capacity to react to them in ways which depend on their internal structure. An activation pattern is "read" by virtue of having effects elsewhere in the system. That is why stability (i.e., a *constant* activation level) is such a crucial feature of activation pattern representations. Being stable enables an activation pattern to contribute to the clamping of inputs to other networks, thus generating further regions of stability, and ultimately contributing to coherent schemes of action. Moreover, the quality of this effect is *structure sensitive* (ceteris paribus), that is, it is dependent on the precise profile of the source activation pattern. While the semantics of a PDP network is not language-like (lacking a combinatorial syntax and semantics) it typically involves some kind of systematic mapping between locations in activation space and the object domain (more on this shortly). And like symbolic representation in digital computers, activation pattern representation involves a one-to-one (or many-to-one) mapping between vehicles and contents; no activation pattern ever represents more than one distinct content. In view of all this, stable activation patterns are best treated as explicit representations.

While activation patterns are a transient feature of PDP systems, a "trained" network has the capacity to generate a whole range of activation patterns, in response to cueing inputs. So a network, in virtue of the magnitudes of its connection weights, and its particular pattern of connectivity, can be said to *store* appropriate responses to input. This long-term storage of information – which sometimes goes by the name of *connection weight representation*[23] – is *superpositional* in nature, since *each* connection weight contributes to the storage of *every* stable activation pattern (every explicit representation) that the network is capable of generating. Consequently, the information that is stored in a PDP network is not encoded in a physically discrete manner. The one appropriately configured network encodes a *set* of contents corresponding to the range of explicit tokens it is disposed to generate. For all these reasons, a PDP network is best understood as storing information in a *potentially explicit* fashion. This information consists of all the data that the network has the capacity to render explicit, given appropriate cueing inputs.

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

NETtalk, PDP and Connectionism

To illustrate some of the features of PDP systems I have just described (and a few others besides) let me turn to a consideration of NETtalk, that darling of philosophical discourse about connectionism. NETtalk is a three layer feedforward network devised by Sejnowski and Rosenberg (1987). Its input layer consists of 203 units arranged into seven groups of 29. Of the units in each input group, 26 encode one letter each of the English alphabet, and the remaining three encode punctuation. It is thus possible to represent a seven character

[23] Strictly speaking, this should be connection *matrix* representation, or some such, since long-term memory in a PDP network supervenes on both connection weights *and* the intra-network pattern of connectivity.

expression over the input array. The input units are fully connected to a hidden layer comprising 80 units, which in turn are connected to 26 output units. The latter are used to code for speech sounds. For a given input string, NETtalk is trained to decide which phoneme corresponds to the fourth character in the string, given the context created by the remaining six. Thus, in order to determine its phonetic transcription, a word must be "fed through" NETtalk's input layer from right to left, so that each letter takes a turn at fourth position. After training (using back-propagation of errors), and when appropriately connected to a speech synthesiser, NETtalk is capable of pronouncing text from its training set with about 95% accuracy. It can handle both regular and irregular cases with equal facility, and is capable of generalising to text not contained in its training corpus. To the extent that it can be characterised as following rules, NETtalk's behaviour innocently emerges from the interaction among its units (mediated by its weights), since none of its representations corresponds to an explicit rule.

At its input layer the representations employed by NETtalk are *local* with respect to letters, but *distributed* with respect to words, since words are represented by patterns of activation over the whole input array. Phonemes are represented in a distributed fashion across the output layer, but the representation here is local when viewed in terms of distinctive speech features (Churchland & Sejnowski 1989, p.30). It appears that the decision to label a scheme of representations 'local' or 'distributed' is, at least in these cases, somewhat interest-relative. Whatever one makes of NETtalk's peripheral layers, it is what happens at the hidden layer that is most striking, and most significant from the perspective of brain-based representation. During training NETtalk discovers a set of weights that encode a large amount of potentially explicit, contextually nuanced data concerning letter-to-sound correspondences in the English language. This information becomes explicit, and is reflected in the activity over the hidden layer, whenever an appropriate string of characters is represented at the input layer.

When one examines the patterns of activation over the hidden units a number of important things emerge. First, for any given instance of a particular letter-to-sound correspondence (of which the English language contains 79) there are typically about 15 units in the hidden layer with significant levels of activation – the remainder showing little or no activity. Thus, the representations in NETtalk's hidden layer are distributed, in the straightforward sense that each input to the net evokes a response involving multiple units.[24] Note, however, that extendedness of representational resources is not particularly significant, in and of itself. As Clark points out (1993, p.19), a scheme of representation in which groups of units are used to encode particular items (the joint activity of units 1, 2 and 3 for item A, units 4 and 5 for item B, and so on) but in which none of the units contributes to more than one representation, is effectively a localist scheme. It is coarse coding – the sensitivity of each unit to more than one type of stimulus condition – that makes distribution interesting, for it ensures that each unit is (diachronically) involved in representing many different items. And it is this sharing of representational resources that gives rise to the superpositional storage characteristics of PDP systems.[25] Given that around 15 of its 80 hidden units are involved in representing each of 79 letter-to-sound correspondences, it is

[24] To be more accurate, NETtalk has *semi-distributed* (as opposed to *fully* distributed) representations over its hidden layer, because not every hidden unit contributes to the representation of each letter-to-sound correspondence.

[25] Explicit coding across multiple units implies that *sets* of weights are involved in storing each distinct response to input, and coarse coding implies that each weight is implicated in multiple responses. When combined these factors imply that information storage across the weights is *superpositional*: all of the stored contents are *simultaneously* encoded by the same (or overlapping) weights.

clear that NETtalk employs coarse coding across its hidden layer. Consequently, NETtalk's capacity to negotiate its target domain is very much dependent on superpositional storage.

Second, it is possible to calculate, for any given letter-to-sound correspondence, the average level of activity (over the relevant portion of the training set) for each hidden unit. This yields a vector expressing the central tendency of the various hidden layer patterns which arise when NETtalk effects a particular letter-to-sound transformation. Repeating this exercise for all the correspondences produces a set of 79 average activity vectors in the eighty dimensional space of hidden unit activations. A fascinating pattern emerges when one measures the distances between these vectors (using a Euclidean metric): there are clusterings among the activity vectors that are very natural given their respective contents. For example, all those vectors associated with vowels (including y when it yields a vowel sound) cluster together in activation space, as do the vectors associated with consonants. Within these groups there are further apt clusterings, such as the letter-to-sound correspondences involving the letter e being close to each other, the correspondences involving the hard c (as in *c*at) forming a tight group, and so on. These clusterings are found to be preserved independently of the network's starting weights, although precise distances in activation space are not preserved. It appears that, as a consequence of learning to produce the required input-output function (the phonemic transcription of a body of text), NETtalk has acquired a set of internal representations with a *semantic metric* (Clark 1993, pp.17-19). What this means is that the natural geometrical measure of similarity among the activation patterns generated over NETtalk's hidden layer (distance in activation space) turns out to be, at one and the same time, a measure of the semantic similarity among the contents carried by those patterns. As van Gelder puts it: "similarities and differences among the items to be represented [are] reflected in similarities and differences among the representations themselves" (1991, p.41). NETtalk is exquisitely sensitive to these differences among its inner vehicles, and is thus able to perform appropriate mappings from hidden unit representations onto phonemes (as represented on its third layer). It treats vectors in one region of activation space as essentially identical, by mapping them all onto the same phoneme, but vectors in other nearby regions as distinct, by virtue of mapping them onto different phonemes. For this reason NETtalk can be said to embody a partition of its hidden unit activation space into 79 distinct *categories* – one for each letter-to-sound correspondence (Churchland 1995, pp.90-1).

Before going on I will briefly mention two further ways in which NETtalk exemplifies the important properties of PDP systems. When a group of letters is represented over NETtalk's input array, the system responds by generating a phoneme appropriate to the letter represented in fourth position. But NETtalk clearly doesn't treat all instances of a given letter as identical; its response is shaped by the particular orthographic context in which that letter is presented. Thus, in the context of a word like 'family', the letter y (which here acts as a vowel) elicits a quite different response to that same letter in the context of a word like 'yet' (where it acts as a consonant). The hidden unit activation patterns in these two cases are actually located in quite different parts of activation space, but, more importantly, the output in the first case is the phoneme /i/ (as in P*e*te), while in the second it is /y/. In other words, NETtalk has the capacity to respond to its input in a highly *context sensitive* manner.[26] One

[26] This way of describing the behaviour of NETtalk raises an important issue about the way hidden unit activation space is typically interpreted. The standard approach, which I have adopted above, is to apply a semantics which is at once *backward-looking* (since it mentions the letter to which a given hidden unit activation pattern is the response) and *forward-looking* (since it also mentions the phoneme onto which that activation pattern is mapped). An alternative is to apply a semantics that is purely backward-looking, and thus, in a sense, faithful to the direction of causation in the network. One achieves this by treating each hidden unit activation pattern as a context-bound representation of a particular letter (the relevant context, of course, being orthographic).

might think that this way of describing things is a bit of a cheat: NETtalk doesn't really know anything about orthographic context, because it has simply been trained to respond appropriately to a *fixed* training set – it is merely a neat realisation of a particular (finite) function. To say this, however, is to miss the significance of the semantic metric that NETtalk evolves during training. By virtue of this representational structure, NETtalk has the capacity to pronounce novel words with reasonable accuracy, i.e., it has the capacity to *generalise*. Unless we grant that NETtalk actually captures something of the abstract structure of English pronunciation, this capacity is just downright mysterious. So NETtalk is more than just a look-up table for its training set.

Both context sensitivity and the capacity for spontaneous generalisation in PDP systems are a direct consequence of employing distributed, superpositional representations. These are but two of the many well advertised features of PDP systems, which also include:

- *Graceful degradation* – the capacity to sustain damage (e.g., the loss of some units) without significant effect on performance, and to deliver sensible responses even when presented with input that is partial, or contains errors.
- *Content-addressability* – the capacity to retrieve complex information based only on access to some part of that information.
- *Automatic prototype extraction* – the discovery, via the learning process, of the statistical centre of a set of training data. In recurrent networks a prototype can take the form of a *point attractor*: a vector in activation space towards which nearby vectors tend to converge during the relaxation process. It is thus possible for a suitably configured recurrent network to treat a disparate set of stimuli as identical under some description.[27]

It is these "unique selling points" (Clark 1993, p.17) which have led a number of theorists to suppose that parallel distributed processing opens up the possibility of a quite radical break with the classical computational theory of mind. Classicism, you will recall, is the claim that human cognition consists in the symbol manipulations of a collection of neurally-realised digital computers (see section 2.2). The PDP computational framework does for connectionism what digital computational theory does for classicism. Human cognitive processes, according to connectionism, 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 behaviour. This is the *connectionist computational theory of mind*.[28, 29]

---

[27] For other lists see McClelland, Rumelhart & Hinton 1986, Clark 1989, Ch.5, and Churchland & Sejnowski 1992, p.171.

[28] Some of the more prominent contemporary philosophers and cognitive scientists who advocate a connectionist conception of cognition include Bechtel (1987), 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).

[29] It is worth emphasising that connectionism is *not* committed to the view that the brain is a single, huge PDP network. Such a view does not accord with what we know about the localisation of function in the brain, nor does it provide connectionism with sufficient flexibility and computational power to account for the kinds of processing and learning of which humans are capable. See Clark & Thornton forthcoming, and Clark & Karmiloff-Smith 1993 for discussion of the limitations of "first-order" connectionism, and some suggestions as to how we might transcend it.

Connectionism and Mental Processing

At this point classicists would be within their rights to ask: What is so special about connectionism? Precisely what does a theory of mind based on the PDP computational framework buy you that can't be purchased with digital coin? In what remains of this chapter I'll address these questions.

In order to distinguish connectionism from classicism a natural place to start is with those features of PDP systems that I've just been at pains to spell out: the use of distributed, coarsely coded (activation pattern) representations; the emergence of a semantic metric as a consequence of training; the capacity for spontaneous generalisation; content addressability; graceful degradation; automatic prototype extraction; and context sensitivity. But the obvious reply to such a list is that these features are either straightforwardly present in digital computers, or can easily be implemented in such systems without undermining their digital status. Content-addressable memory, for example, can be implemented in a digital computer using some kind of indexing scheme – a list, for each possible content, of the memories which contain that content (McClelland, Rumelhart & Hinton 1986, p.26). Robust performance in the face of damage can be achieved simply by building a certain amount of redundancy into the system, for example by "cloning" representations. This solution relies on distributing the representational load in such a way that no single computational element is crucial to successful performance. As remarked earlier, mere extendedness of representational resources, in and of itself, is not a particularly significant feature of PDP systems, and distribution (in this sense) clearly doesn't mark a useful contrast between classicism and connectionism. One might imagine that insistence on coarse coding – the involvement of each computational element in the representation of more than one content – would give rise to a more significant contrast. However, as van Gelder points out: "when numbers are encoded as strings of bits in a register of an ordinary pocket calculator, they are being represented by distinctive activity patterns, and each unit or location participates in the representing of many different numbers over the course of a calculation" (1991, p.36). Classicism can thus avail itself of coarsely coded, distributed representations, without fear of contradiction. Similar remarks apply to all the items in connectionism's glory-box.[30]

A common connectionist reaction to these claims is to point out that, while it may well be possible for digital systems to exhibit many of the properties listed above, they can only do so by dint of complicated extensions to their basic modes of information storage and processing (see, e.g., McClelland, Rumelhart & Hinton 1986, p.31). PDP systems, on the other hand, give rise to content addressability, spontaneous generalisation, error tolerance,

---

[30] Clark argues (1993, pp.23-33) that the context sensitivity of PDP systems is "intrinsic", and hence distinctive of connectionism, because, unlike the kind of context sensitivity that is possible in classical systems, it doesn't rely on a set of transportable, context-free representational atoms. In classical systems, occurrent representations are constructed from a common pool of stored primitives. Thus, semantically related representations with complex contents will, in general, have semantically significant parts in common (this is a consequence of employing a combinatorial syntax and semantics). By contrast, (internal) activation pattern representations are created on-the-fly, using only local resources, as a context-sensitive response to current input. On different occasions even a single item may be represented by distinct vehicles; vehicles which resist decomposition into interpretable parts, but whose structural differences reflect computationally significant differences in context (see Elman 1990). To identify context-free symbolic primitives in a PDP system one would need to "discover syntactic structures which *persist* unaltered and carry a fixed content" (Clark 1993, p.32). However:

> The only contenders for such persistent content bearers are the weights which moderate the various context-sensitive activity patterns. But a given weight, or set of weights, cannot be identified with any fixed content in these superpositional systems, for each weight contributes to multiple representational abilities. (1993, p.32).

Superpositional storage thus appears to ground the distinctive kind of context sensitivity evident in PDP systems. At root this amounts to a difference in the *style of processing* employed by these systems – a point I will develop in what follows.

and so forth, as natural by-products of the kinds of representation and processing they employ. Given the significant place of such attributes in our cognitive profile, it is fair to say that a good deal of biological and evolutionary plausibility accrues to connectionism on this account (Clark 1989, Chs.4 & 5). What is it that allows PDP systems to realise these attributes so naturally and efficiently? The short answer is "superpositional storage", but to answer in this way runs the risk of obscuring what it is that really divides digital and PDP systems. And it risks ignoring the most significant promise held out by connectionism. A further brief look at the nature of representation in PDP systems, through the clarifying lens of Dennett's taxonomy of representational styles, will help to avoid this outcome.

I showed above how Dennett's taxonomy can be used to identify non-classical versions of explicit, potentially explicit and tacit representation in PDP systems. However, what has become apparent in our subsequent discussion is that the distinction between potentially explicit and tacit representation actually lapses for connectionism. 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 the network's connection weights and its pattern of connectivity. But, as we saw above, these connection weights and connectivity structure also encode a network's tacit knowledge, because they are jointly responsible for the manner in which it processes information. Tacitly represented information thus has the *same* supervenience base as potentially explicit information, in a PDP system. In effect, 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.[31]

The integration of potentially explicit and tacit representation lies at the heart of PDP. In more conventional terms it amounts to a break with the code/process divide: the separation of memory and processing that is characteristic of digital systems. Digital computers store at least some data in an explicit form. By contrast, PDP systems store *all* of their information in a potentially explicit fashion. Consequently, given 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 in which it is *everything* a network knows – the network's "memory" – that actually governs its computational operations. This deep interpenetration of representation and process has important consequences for connectionism. Not only does it account for the efficiency with which PDP systems realise many of their important emergent properties (content addressability, automatic prototype extraction, etc.), but it suggests a potential solution to one of the most pressing problems facing cognitive science at this time: the need to explain the fast, flexible,

---

[31] One might think that there is still a way of maintaining a distinction between potentially explicit and tacit representation within the PDP framework. Tacit knowledge has frequently been identified with rule-following (as in tacit knowledge of English grammar). PDP networks can also be regarded as following rules. For example, NETtalk can be said to have tacit knowledge of the rules of English grapheme-to-phoneme transformation. It is important to realise, however, that in saying this one is adopting a *diachronic* view of network behaviour. Rule-following can't be discerned in single input-output pairs – it only emerges when we examine semantically related sets of transformations, each of which must be tokened on a separate occasion. Thus, a network's tacit knowledge can, perhaps, be equated with the rules that are dispositionally encoded in its weights, and diachronically realised in its behaviour, leaving potentially explicit representation to line up with the individual transformation instances. Strictly speaking, of course, there is nothing more to the "rule-following" of a network than the set of individual input-output pairs it has the capacity to produce. Having said this, I've no real objection to taking the long view of network behaviour, primarily because it fails to undermine either of my central claims, namely that: (1) the supervenience bases of tacit and potentially explicit representation in PDP systems are identical; and (2) everything a PDP network knows, whether it be tacit, or potentially explicit, is simultaneously brought to bear in determining processing outcomes.

and profoundly global nature of human cognition. Let me finish the chapter by examining this issue.

The Systematicity and Global Sensitivity of Thought

Cognitive science is in the business of explaining intelligent behaviour. The most fundamental methodological precept of this discipline is that intelligence ought to be accounted for in terms of semantically coherent causal processes defined over in-the-head representations. We have seen that there are currently two different suggestions as to the nature of those causal processes, and thus two distinct computational theories of mind on offer. What should constrain our choice between them? Clearly, evolutionary and neurobiological plausibility, and congruence with what we know about the task demands of both perception and cognition, should be near the top of our list. Two constraints are of particular significance. The first is the capacity to explain the "combinatorial" properties of thought: its productivity and systematicity. As I indicated in section 2.2, classicism has a nice neat story to tell about these aspects of thought. Connectionism, on the other hand, has been accused of failing to satisfactorily account for both the productivity and the systematicity of thought. In discussions of the relative merits of classicism and connectionism the systematicity issue, in particular, has been very prominent.[32]

Somewhat surprisingly, a second, and in many ways more significant constraint, has received far less discussion in this context. I have mind here the capacity to explain the global sensitivity of cognition. This sensitivity emerges in a number of ways, but primarily in our ability to:

1) update our stored knowledge-base as new data arrives, in a way that is sensitive to the relevance of this data to current, and ongoing projects, and to its possible consequences for other stored information; and
2) retrieve and utilise the stored information that is most relevant to current projects.

These two abilities constitute the core of intelligence, because our capacity to respond appropriately to novel or changing environmental conditions is built on a foundation of sensible learning (learning what we need to learn), and contextually nuanced thought. What makes knowledge update globally sensitive, is the fact that practically anything one knows could potentially be affected by incoming data, and practically anything in that data could be relevant to one's goals. What makes thought global is that almost anything one knows could turn out to be relevant to achieving one's current goals, and thus may need to influence a thought process directed at those goals. As Fodor puts it: "there seems to be no way to delimit the sorts of informational resources which may affect, of be affected by, central processes of problem-solving" (1983, p.112).

Classicism has a good deal of difficulty in accounting for this global sensitivity, because the only way a classicist can allow stored information to affect the course of processing is for it to enter the causal chain in the form of an explicit token – in Fodor's memorable terms: "No Intentional Causation without Explicit Representation" (1987, p.25) What makes this problematic is that in order to determine which of one's current beliefs need to be updated in the light of incoming data, for example, a digital system must retrieve and token them all, there being no way to determine in advance which of them will be relevant. And this kind of

[32] See Fodor & Pylyshyn 1988, Fodor & McLaughlin 1990, and Fodor 1997.

search leads to familiar problems of combinatorial explosion, rendering it computationally intractable (at least, so far as real world, real-time problem-solving is concerned).[33]

Connectionism, by comparison, seems well placed to deal with the global sensitivity of thought. What works in connectionism's favour is superpositional storage, and the tight relationship between representation and processing that is characteristic of PDP systems. Whereas a digital computer must explicitly token all the information that it processes, a PDP system, by virtue of storing its data in the very substrate that supports processing, need not render stored contents explicit (in the form of stable patterns of activation) in order for those contents to affect cognition. A connectionist is thus free to suppose that a globally sensitive task – a task in which a great deal of what we know must be simultaneously brought to bear – simply involves a process of stabilisation taking place over some suitably large region of cortex, whereby a number of networks simultaneously settle into stable states. While this process is in train the only explicit representations that need exist are those that originate with sensory input, and some further relatively localised representations[34] that act as the locus of subsequent activity. Processing is governed by the total pool of information superpositionally encoded in the relevant portion of the cortex, and completes when the system as a whole settles down into a global state of stable activation.[35] It is only at this point that any explicit information is generated. This style of information processing has the advantage of being very rapid, even though it is subject to the influence of large bodies of stored information, because such information doesn't have to become explicit to throw its weight around. Knowledge update is also amenable to this kind of treatment: superpositional storage ensures that changes to one part of the knowledge base, via processes of synaptic modification, will automatically flow on to related contents, given that they are stored in the very same weights.

The irony of all this, as Clark points out (1993, pp.108-11), is that classicists tend to reject connectionism on the grounds that it (allegedly) fails to account for the productivity and systematicity of thought, despite the fact that classicism itself has had so little success explaining an equally deep feature of thought, namely, its global sensitivity. And, as I have demonstrated, it is in just this regard that connectionism offers some real hope of progress. The dispute between classicism and connectionism is thus something of a stand-off. Each theory of mind has its own pattern of strengths and weaknesses, and neither convincingly dominates the explanatory territory at this stage.

---

[33] For discussion see, e.g., Dennett 1984, Haugeland 1987, and Copeland 1993, Ch.5.

[34] These explicit representations (which are often largely linguistic, I suspect) provide the stable input required to initiate processing elsewhere in the system. On many occasions this input is probably supplied by the sensory systems alone.

[35] It is worth noting that the process of activation passing and stabilisation is self-limiting, because any network which is not sensitive to the kinds of influences currently passing through the system will necessarily fail to stabilise. For this reason, only information that is relevant to the task at hand has any chance of being rendered explicit.

3

Phenomenal Experience: A Brief Survey

Of all the concepts we inherit from the folk, none is currently more subject to confusion and general disarray than consciousness. And there is no phenomenon that is less apt to generate theoretical consensus than this same feature of mind. Therefore, before hazarding a connectionist theory of phenomenal experience a scrupulous examination of the explanandum is in order. In this chapter I will attempt to provide a straightforward, intuitively plausible characterisation of phenomenal experience. This characterisation will act as a benchmark for the connectionist theory of consciousness to follow (in Chapter 7). I will also address the relationship between consciousness and thought, since it has a bearing on a number of issues to be addressed in subsequent chapters.

3.1 Some Forms of Conscious Experience

A theory of consciousness, like any other theory, must begin by isolating its domain in some way. That is, theorists must have some way of drawing a boundary around the phenomena they seek to explain. Of course, as theory building commences the nature of the target may come to be reconceived in certain respects (as, for example, when parts of a domain are eliminated), but this does not remove the initial requirement to carefully specify the explanandum. When it comes to phenomenal experience this task is unusually difficult, because, unlike other domains of investigation, the starting point for theories of consciousness is ineliminably subjective. Whereas a theory of chemical interaction, for example, can point to various publicly observable phenomena that the theory must account for (e.g., at the grossest level, the various colour changes, gaseous emissions, changes in weight and density, etc., of substances under combination), a theory of consciousness can only ask each of us to turn inwards and attend to our own experience. Thus, we are faced with a special epistemological problem in relation to building a theory of phenomenal experience. Some have regarded this difficulty as so acute that it ought to exclude consciousness from scientific treatment altogether, and have spent much energy advertising the pitfalls of introspection and the dangers of the unobservable. However, in this century our experience with theories of both the very large (cosmology), and the very small (quantum theory), has made us less wary of the unobservable. And when it comes to the restricted access that consciousness affords, the least we can say is that conscious experience, although in some sense private, is, for each of us, surely the most salient feature of reality, since it is the ground and condition for all observation.[36] I propose, therefore, to assume, without further argument, that conscious experience is a suitable object of scientific investigation.

What, then, is in need of explanation? In answer to this question the best one can do is ostend. Chris Mortensen, a philosopher at the University of Adelaide, has for many years identified the relevant target by getting his students to follow these instructions: "close your

[36] The rise of cognitive science has been instrumental in reorienting our attitudes towards consciousness. However, some of the old fears are still with us, particularly in relation to introspection. I will have more to say about this issue, and the way its importance has been exaggerated, in what follows.

eyes; now open your eyes; now close them again." He then remarks: "The thing that came into existence between the two occasions of eye-closing – that was a visual experience. It is that kind of thing we are referring to when we speak of conscious experience." Apart from the occasional disciple of Berkeley in the audience (who might suppose it was the room that came into existence), people usually get the point pretty quickly. And once you notice phenomenal experience, i.e., once you are ready to move beyond a naive realism, and make the distinction between appearance and reality, then you discover that conscious phenomena are both rich and ubiquitous. They are ubiquitous because they are the constant condition of our waking existence (not to mention some of our sleeping existence), and they are rich because they are multi-modal and complex.

To begin with, phenomenal experience comes in at least the following distinct sensory varieties: visual, auditory, olfactory (smell), gustatory (taste), tactile, thermoreceptive (associated with sensations of heat and cold), proprioceptive (associated with awareness of the activities of muscles, tendons, and joints and of equilibrium), and the various kinds of pain. We can distinguish these in experience – each of these modalities has a distinctive "feel" – and we also know something about their particular receptive and neural bases. Neurophysiologists divide the sense modalities into two broad classes: the *general* and the *special*. Each modality has specialised receptors (which respond to a particular kind or kinds of environmental stimuli), a particular post-stimulus neural pathway, and distinct sites in the brain to which these pathways project. The general senses – the varieties of touch (light touch, pressure, vibration), temperature sensitivity, proprioception, and pain – are characterised by having structurally simple receptors that are both widespread and numerous. For example, pain receptors (called nociceptors) are simply the branching ends of the dendrites of certain sensory neurons, and are found in practically every tissue of the body. On the other hand, the special senses – vision, audition, taste and smell – have complex receptors that are found in only one or two specific regions of the body. Another important difference between the special and general senses is that, in the case of the latter, there is a common region of the cerebral cortex into which the various kinds of receptors initially project (the somatic sensory or primary somesthetic area), whereas each of the special senses has its own anatomically distinct primary site. The significance of this is that the resources devoted to the special senses (just in terms of gross numbers of neurons) are clearly greater.

While on this topic, it is worth noting the distinction between receptors that respond to stimuli originating outside the body (known as *exteroceptors*), which are located near the surface of the body, and give rise to sensations of hearing, sight, smell, taste, touch, pressure, temperature and pain, and those that provide information about the internal (somatic) environment. The latter include the *visceroceptors*, which give rise to sensations of pain, pressure, fatigue, hunger, thirst and nausea (all originating from within the body), and the *proprioceptors*, which provide information about body position and movement. Notice that this distinction (between extero- and entero- receptors) cuts across the distinction between general and special senses, but in many ways it is more significant, for our purposes, because it provides an anatomical basis for distinguishing between experiences that relate to the external world, and those that relate to the body.[37] A further significant divide within

[37] Sadly, receptor location isn't a perfect basis for this task. For example, it is notorious that sensations of heat and cold are an unreliable guide to the temperature of objects impinging on the body (see Churchland 1979, for discussion), so it is unclear whether the experience here concerns the temperature of the skin (and so is strictly speaking an instance of enteroception), or the temperature of the object (and thus constitutes exteroception). And when one determines the shape of an object through touch, is this an instance of pure exteroception, or is our phenomenology perhaps dual, containing an experience of the body surface, which is then used to generate a perceptual content relating to the external object? Nevertheless, for all that, the location of receptors is a fairly reliable guide to the kinds of information which the associated experiences convey.

the receptive basis of sensation concerns the varying capacity of receptors to adapt to ongoing stimulation. Rapidly adapting (*phasic*) receptors (associated with touch and smell), quickly lose their sensitivity to stimuli, and are thus suited to alerting us about changing conditions. Slowly adapting (*tonic*) receptors, on the other hand (associated with, e.g., pressure, pain and body position), are suited to providing information about steady states of the body. The obvious connection with conscious experience here is that some experiences are short-lived, even in the continued presence of the stimulus source (e.g., a strong smell quickly diminishes in intensity even when the source of the smell is not removed), while others are coextensive with the stimulus duration.[38]

One might wonder what relevance all this neurophysiology and anatomy has for an examination of conscious experience. I suggest that it has a number of things to teach us. Since I am here assuming physicalism, it goes without saying that a structural and functional analysis of the nervous system is likely to pay some dividends towards an analysis of experience. At the least it can act as a taxonomic aid, directing our attention to easily overlooked features of consciousness (for example, it is easy to leave proprioceptive and thermoreceptive sensations off the list of phenomenal experiences). We can see, in fact, that just like its neurological basis, sensory experience is highly structured. It incorporates a diversity of distinct modalities, each of which appears to be subserved by some relatively independent neurological resources. And the extent of these resources, in each case, appears to be correlated with the comparative richness of the phenomenology in that modality. In addition, we've seen that there is a connection between the temporal characteristics of sensory receptors and the relative latencies of various kinds of experience, and that there's some anatomical basis for distinguishing experiences concerning the internal environment and those that embody information about the outside world.

So conscious experience comes in a variety of modes, which differ in terms of their relative latencies, their comparative richness, and their phenomenal "feel". Furthermore, it is crucial to recognise that moment by moment experience typically encompasses several of these modes *simultaneously*. If you need proof of this, just notice that as you read this page your visual phenomenology is accompanied by concurrent auditory experiences (perhaps the sound of distant traffic, or people talking) and proprioceptive experiences (the arrangement of your limbs; the feeling of the chair against your body). Neither of these disappears entirely when you attend to visual experience, even if there is a heightening of perception in relation to the object of focal attention. Not only are the various modes distinguishable within instantaneous consciousness, but it appears that any one of them can be removed or lost without affecting the others (try closing your eyes for a moment). In other words, consciousness is an *aggregate* – a sum of relatively independent parts. These parts are like so many strands in a woven cloth – each strand adds to the cloth, but, since they run side by side, the loss of any one strand doesn't deform or diminish the others, it merely reduces the total area of fabric. What I'm suggesting is that at any moment in time phenomenal experience comprises contents from a number of distinct modalities, and that these modalities form independent *parallel streams* of awareness.[39]

This simple analysis of consciousness (into its various modes) only skims the surface as far as the richness of phenomenology is concerned. When we turn to any particular modality, we find that it is bursting with internal structure. Visual experience, for example, is composed of sensations of colour, form, and motion. Recent work in the neurosciences supports this analysis of visual experience, since it shows visual processing to be highly

---

[38] This material was largely gleaned from Tortora & Anagnostakos 1987.

[39] I will have more to say on these issues shortly.

modularised; the visual cortex appears to contain separate subsystems for the processing of information corresponding to each of these distinct kinds of sensation (see Zeki 1993). And just as the various modes of experience seem to be quite independent of one another, there also appears to be a degree of independence *within* modalities. Numerous deficit studies suggest that, as a result of injury to specific sites in the visual cortex, any of colour, form, and motion can be lost, *without significantly affecting the others*.

Take the perception of motion for example. Zeki relates the case of a woman who, due to a vascular disorder in the brain which resulted in a lesion to a part of the cortex outside the primary visual area, lost the ability to detect motion visually. This was so severe that,

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

Zeki notes that this was not a total defect in the appreciation of movement "because the perception of movement elicited by auditory or tactile stimulation was unaffected" (ibid.). Moreover, her perception of other visual attributes appeared to be normal. Similarly striking case studies are available in relation to the loss of colour sensations (see, for example, Sacks 1995, pp.1-38). So even within modalities, it appears that conscious experience has components, which are phenomenally distinct, and which can come apart (i.e., can exist even in the absence of some of the others). The parallel, composite nature of phenomenal experience appears to be both an *inter*-modal and an *intra*-modal affair.

Abstract Experience

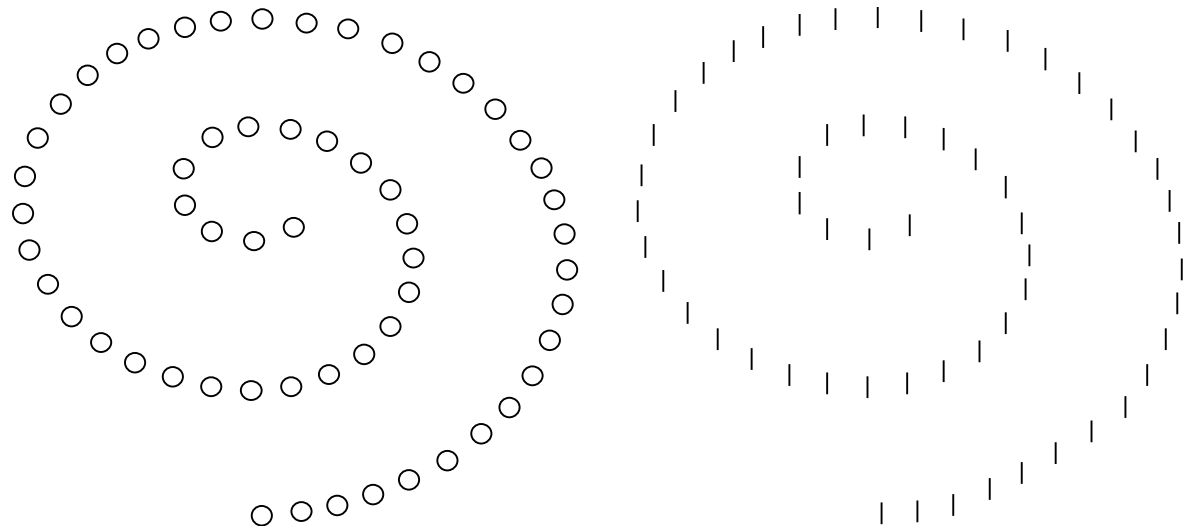
Apart from the basic features I've just described, visual experience also contains a good deal of what is best described as *abstract* phenomenology. This division (into more concrete and more abstract) is a general feature of conscious experience, but I will begin by developing some examples from vision. Consider first the perception of depth. The capacity to perceive the relative distances of objects depends on the fact that we have two eyes, and seems to involve the detection of small disparities between the relative positions of objects in the right and left visual fields. This capacity is demonstrably not due to an inference based on past experience of the relative sizes of objects, because our visual system generates sensations of depth even when presented with a pair of random dot images in which one of the images contains a region that is displaced relative to the other (known as a random dot stereogram). In order to create what is, in this case, the illusion of depth, the visual system must be relying on quite low level cues, since there are *no* objects discernible in either of the separate images.

We can only perceive relative distance out to about one hundred metres, because the further objects are from us, the smaller the disparities between the two visual fields, and beyond this distance the disparities become too small for depth perception to be maintained. An interesting experiment in phenomenology is to stand at the window of a three of four storey building, and attend to the sensation of depth as one first looks at some nearby objects (a stand of trees, for example), and then moves one's gaze outwards. If there is a sufficient range of objects in view it is quite striking that from about one hundred metres and beyond sensations of depth disappear entirely – a quality of flatness pervades the view.[40] One can

[40] Of course, we do still have the abillity to *infer* relative distances; what disappears beyond a hundred metres is the immediate sensory *experience* of depth. For a far more detailed discussion of human stereopsis see Churchland 1995, pp.57-79.

get this latter effect much more easily, of course, by simply shutting one eye. Most scenes then lose something, a quality of extension, which returns immediately upon opening the closed eye (try this with a set of exposed beams in a ceiling, or a row of books along a bookshelf). The important point here is that depth perception is something added to basic visual experience – one can have rich and informative visual experience without it – yet it is a genuine part of the phenomenology when it is present (there is "something it is like" to perceive depth).

What essentially makes sensations of depth abstract is their reliance on earlier and more basic processing. Processing of the dual inputs (one from each eye) to the left and right



Spiral A Spiral B

Figure 3.1 Example of a visual gestalt.

visual fields is necessary before the disparities required for depth perception can be computed. So in some sense depth perception arises further down the processing chain than, say, the early processing associated with the individual retinas. Of course, information concerning depth is not *very* abstract (in the specified sense), it is quite a low-order visual feature, subject to early detection under normal circumstances. For a more abstract feature of perception we must turn to object recognition.

Visual experience consists not only of lines, colours and regions, but of organised wholes. An earlier tradition in psychology called these *gestalts*. For example, Spiral A in Figure 3.1 has recognisable parts, but, in addition to the recognition of these parts, our experience of it contains a distinct phenomenal element corresponding to the whole. This extra element of experience, this quality of "spiralness", is not to be confused with the experience of the separate parts. It is, rather, an experience of the parts in relation to each other – the perception of an organised whole.[41] That the whole has its own distinct phenomenology can be discovered by considering Spiral B. Even though composed of quite different parts it has the same abstract quality we associate with Spiral A. Moreover, like most aspects of visual experience, this quality can be lost, even while the capacity to

[41] Another nice example comes from audition: when one recognises a tune one is also experiencing a gestalt, because to hear a tune is to perceive a relation between the notes of which it is composed, not just the notes themselves; if not, it is hard to explain our ability to recognise the same tune in different keys.

recognise the parts is spared (see the discussion of a patient suffering from carbon monoxide poisoning in Zeki 1993, pp.276-8).

If you are not convinced by this kind of case, consider Figures 3.2 and 3.3. These are well known ambiguous figures; images capable of more than one high-level interpretation. Figure 3.2 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 as a flat line drawing, with no sense of perspective. And whichever of these interpretations one adopts, the details of line and space remain the same. Similarly with Figure 3.3. Whether one interprets it as a vase (dark figure, light background), or as a pair of faces (light figure, dark background), there is no change in the experience of tone and line itself. Thus, both of these figures generate some primary visual experience (i.e., the experience of lines, boundaries, light and dark regions), to which an element of abstract phenomenology can be added. What is striking in this instance is that there is a degree of choice as to the nature of the additional experience. Again this abstract phenomenology is somewhat dependent on lower-order processing, but is neither compulsory, nor entirely fixed.

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

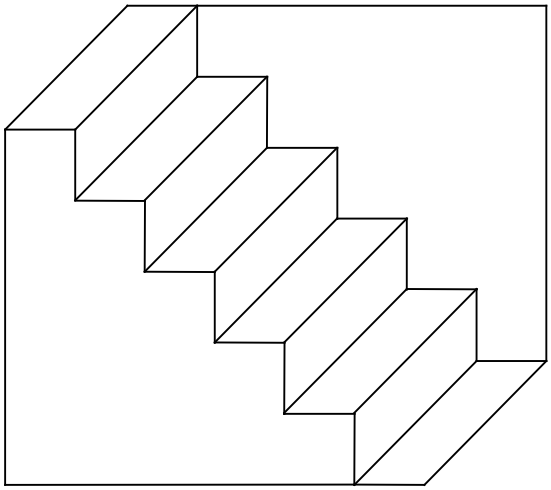


Figure 3.2 Inverting stairs ambiguous figure.

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 recognise a familiar face. Note that this case is slightly different from the gestalt experiences described above, because I am here describing an element of experience further to the mere perception of a face as an organised whole. A familiar face is not only perceived as a face, but as a face with a familiar "feel". This feeling of familiarity (also known as a "feeling of knowing", see Mangan 1993b) is superordinate to facial perception *simpliciter*. It is also to be distinguished from the capacity to associate a name with a face. For those, like myself, 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.

It appears that quite a large region of the cortex (on the underside of the temporal and occipital lobes on both sides of the cortex – see Geschwind 1990) is devoted to high-level perceptual tasks connected with the human face. Damage to this region can result in a number of perceptual deficits, the most striking of which is prosopagnosia: the inability to recognise familiar faces. Prosopagnosics are frequently unable to recognise 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 recognise yourself in a mirror?", he replied, "Well, I can certainly see a face,

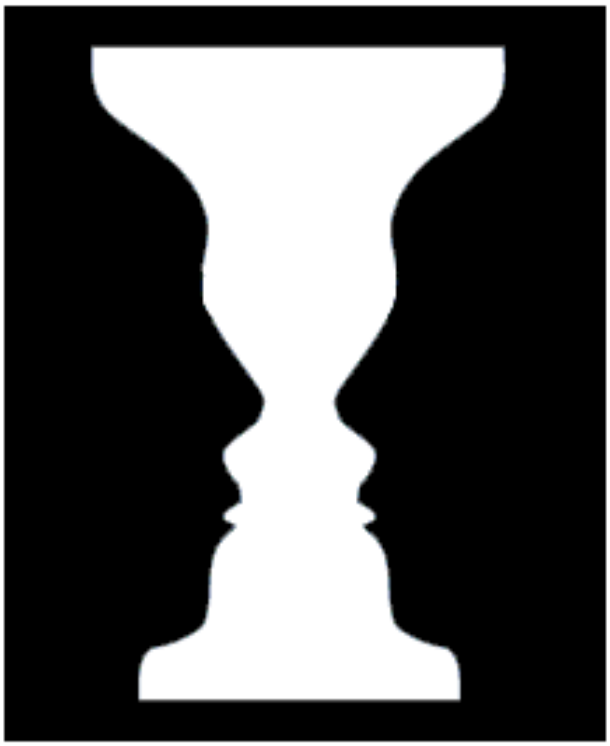


Figure 3.3 Vase/faces ambiguous figure.

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 capacity to experience a face and its various components. Moreover, the lesions responsible for this deficit can spare the capacity to recognise the expression on a face, even when the face itself is unfamiliar. So the recognition of familiar faces and the recognition of facial expression appear to be governed by different cortical regions, and to be quite independent of one another. All of this adds weight to the suggestion that facial familiarity is an independent, abstract element of perceptual experience, which can be effaced without loss to the other elements of visual experience.

As a final example of abstract phenomenology, and one that arises outside the visual modality, consider the well known experience of having a name "on-the-tip-of-the-tongue". This experience can last for some time, with a tantalising ghost or outline of the name coming in and out of consciousness. William James, a famous chronicler of conscious experience, describes it this way:

> Suppose we try to recall a forgotten name. The state of our consciousness is peculiar. There is a gap therein; but no mere gap. It is a gap that is intensely active. A sort of a wraith of a name is in it, beckoning us in a given direction, making us at moments tingle with the sense of closeness, and then letting us sink

> back without the longed-for term. If wrong names are proposed to us, this singular gap acts immediately so as to negate them. They do not fit into its mould. (1890, p.251)

The process of recall is, as James so vividly describes, quite rich in phenomenology. It is reasonable to suppose that tip-of-the-tongue phenomena exhibit abstract experience in its pure form. This "wraith of a name" James refers to, is perhaps an abstract word-gestalt, the form of a word, which for some reason has failed to be filled in with concrete perceptual details. It is to the normal experience of a word, as the quality of spiralness is to any particular, detailed experience of a spiral. No two words we are seeking are quite the same with respect to this form-quality, which is why we can reject wrong guesses so easily. The active "gap" itself can vary; on one occasion "the rhythm of [the] lost word may be there without a sound to clothe it", on another "the evanescent sense of something which is the initial vowel or consonant may mock us fitfully, without growing more distinct" (ibid.).

All these examples of abstract experience, i.e., depth perception, object gestalts, facial familiarity, and so on, have in common a certain looseness of fit between the abstract percept and its underlying, more concrete ground. This is most strikingly evident in the various deficits that arise due to localised cortical lesions, but even in undamaged brains we've seen that early processing does not completely constrain higher-level components of experience. Depth perception can be subtracted from visual experience merely by closing one eye, and object gestalts (especially in relation to ambiguous figures) seem to be subject to a degree of choice. This adds weight to the claim that phenomenal experience is not an undifferentiated whole, but a complex *composite* of many elements; elements that are distinguishable within consciousness.

The discussion to this point has only scratched the surface as far as the richness and complexity of consciousness is concerned. Among humans there is a realm of experience associated with our powers of thought, and our linguistic abilities in particular, that hasn't yet been touched on. I'll take up the task of mapping some of this terrain in the next section.

3.2 Consciousness and Thought

I claimed in the previous chapter that cognitive scientists are united in their understanding of thought to this degree: that they all conceive it in computational terms – as semantically coherent causal operations defined over inner, content-bearing states. This conception is neutral both as to the grain of contents and the computational architecture required to account for human intelligence. You will recall that there is also general consensus regarding our experience of thought: thought processes are largely unconscious (see the introduction). There is something to be said for this. While I am conscious of the moment at which I reach a conclusion, or make a decision, it is not always clear how I got there – a substantial portion of the decision making process is quite inaccessible. Simple thoughts are the most compelling cases. If I ask you to recall your address, or multiply two small numbers, the result normally arrives quickly, but it is hard to give any account of the mental operations involved. Or consider this recipe for finding the right piece in a jigsaw-puzzle: scan the available pieces until one of them grabs your attention. This usually works, but the recognition process it relies on is quite opaque. It is clear that any account of phenomenal experience must ground what is conscious in a great deal of unconscious activity.

What, then, is the relationship between phenomenal experience and thought? The short answer is that conscious states are, in some sense, the *products* of thought processes. Even the stages in an identifiable sequence of conscious states (such as the steps in mental arithmetic) are intercalated with unconscious activity. Consciousness is like the exposed surface of a buried rock, or the cone of a volcano – it is the visible result of a largely subterranean

dynamic. Be that as it may, conscious experience is generated by all kinds of thought processes, from the most mundane, to the most abstract. For example, in addition to the meanings of the words and sentences I'm now typing, I'm also conscious of the shapes of the letters involved, and of the feel of the keys as I strike them. Even on the most generous construal of the unconscious, it is undeniable that thought generates a large array of conscious contents. The material in Section 3.1 was an attempt to map some of that terrain.

However, all of the examples I discussed above (such as depth perception, and facial recognition), while dependent on some degree of hard-wired intelligence, are the products of relatively simple thought processes; processes that are rapid, mandatory and cognitively impenetrable (see Fodor 1983, pp.52-86 and Pylyshyn 1980). Creatures that rely entirely on such processes (interacting with an appropriately configured motor system) in order to determine their behaviour, still qualify as thinkers, on the generic computational theory of mind, since this only requires that their intelligence, such as it is, be dependent on operations that implicate informational contents. But human intelligence is special, at least in degree, if not in kind. When we speak of 'thought' in relation to humankind, we normally have in mind something other than the operation of systems that lead fairly directly to a motor response. Human thought, particularly when it involves symbol systems of one kind or another, is relatively slow, and involves complex, internally-driven sequences of states (think of mental arithmetic again). Therefore, in order to pursue the relationship between consciousness and *human* thought we need to be able to identify various levels or kinds of thought. In other words, we need to be able to mark the contrast between what I'll style *higher-thought* and those phylogenetically older kinds of thought with which we began.

Higher-thought

It is reasonable to suppose (and supported by the available evidence) that the internal processes responsible for behavioural plasticity, in all its many forms, come in a corresponding variety of forms and degrees of complexity. Reflexive behaviours of various kinds (such as the withdrawal reflex in response to contact with a sharp object), which constitute a small degree of hard-wired intelligence, are known to be supported by short and simple neuronal arcs. From here we move to behaviours (such as avoiding a fast moving object), which, although they implicate complex representations of the world, are still governed by relatively direct pathways from sensory regions of the brain to motor control centres. Beyond this are temporally extended behaviours, such as playing a piano, or a round of golf. These exhibit a degree of automatism (one can learn to initiate sequences that require minimal conscious control), but nevertheless suggest the existence of complex, nested control structures, subserving a hierarchy of goals.

While simple reflexes can operate in the absence of any conscious experience, not so the other behaviours I've mentioned. One can't avoid a fast moving object unless one can see it (although admittedly the phenomenology here may be peripheral and fleeting), and temporally extended behaviours clearly implicate a good deal of conscious experience, including consciousness of goals. Indeed, the further up the scale we ascend, the more significant the role of consciousness becomes.

Another significant indicator of the ascent from lower to higher thought is the changing role and importance of environmental input. Reflexive behaviours are very much input-driven, and usually can't be avoided. Playing the piano, on the other hand, despite the ongoing sensory feedback required for its maintenance, is an internally initiated activity. Pen and paper arithmetic shows a similar mixture of ongoing environmental feedback, in the form of the changing marks on the paper (which are acting as a form of external

memory), and internal control.[42] In both cases, however, the internal processing is relatively direct. To determine that six sevens are forty-two, for example, one simply needs to become conscious of the problem to be solved, and the answer will (usually) present itself. But pen and paper arithmetic can be internalised, and when one engages in *mental* arithmetic one can discern *sequences* of conscious states between input and output, as I suggested above. Moreover, once such activity is initiated, it is capable of being maintained with little or no environmental input, over relatively lengthy periods of time.

There are two morals to be drawn so far regarding the nature of higher-thought: it is characterised by *extended sequences of conscious states*; and it is largely *internally driven*. Clearly these indicators do not mark a sharp dividing line between lower and higher thought. There is a continuum of internal processes, ranging from those that are highly environment dependent, and exhibit little or no conscious involvement, through to those that are largely self-initiated and maintained, and contain long sequences of conscious intermediaries between input and output. Nevertheless, the processes we are particularly interested in are those at the far end of the scale, since it is these that are most characteristic of human thought.

But we can do better than this. So far I have provided only the grossest kind of characterisation both of our behavioural repertoire, and of the internal processes responsible for its maintenance. Cognitive psychology has for some time been developing a much richer account of thought, albeit a rather disunified one that has sometimes tended to confuse the task and algorithmic levels of description. Philip Johnson-Laird (1988) has undertaken to tighten up this account in the form of a taxonomy of thinking, which it will be useful to present here.

Johnson-Laird's taxonomy attempts to fit the domains of thought into a single framework specified at the task level, i.e., at what David Marr (1982) calls the "computational" level: the level of description where one specifies what the mind is doing, rather than how it is doing it. The basis of the taxonomy is a series of questions that generates a tree-structure, with the various kinds of thinking as the leaves. Johnson-Laird first asks: Does the thinking have a goal? If the answer is no then the thought process is what he calls *association*, the free, and seemingly random interplay of ideas characteristic of dreams and daydreaming. If yes, then the thinking is some species of problem solving. The first division within this family is generated by the question: Is the thinking deterministic? By this Johnson-Laird means: does it follow a path towards its goal such that at every stage the next step is completely determined by the current state of the process? (This is by contrast with non-deterministic computational processes, which can yield more than one output from a given input and internal state.)[43] If the answer is yes then the thought process is some form of *calculation*. Mental arithmetic is the simplest example here. Such processes have a single, well defined goal, towards which they move in a lock-step fashion. If no, then a further question arises, namely: is there a precise starting point to the process?

A positive answer to this question brings us to *induction* and *deduction*, which differ according to whether they increase semantic content or not (induction does, deduction doesn't). Both these kinds of problem solving start from some kind of representation of the world and move through the problem space in a non-deterministic way, i.e., at each point there is more than one possible way forward. If one starts from a linguistic input (a set of premises), then induction arrives at its goal by deriving a conclusion that is consistent with, but goes beyond, the premises. This is the kind of activity that Sherlock Holmes was

[42] Unlike the former case, the feedback here isn't governed by any significant temporal constraints.

[43] Another way of putting this would be to ask: Is there an algorithm for the specified goal?

famously engaged in.[44] The conclusion of an induction contains more information than strictly follows from the premises. Deduction, on the other hand, leads to a conclusion that is necessarily true if the premises are true, and therefore does not increase information content. It is important that induction and deduction be understood here in the widest possible sense, i.e., they should extend to physical problem solving, in which natural language may have little or no role. In such cases the premises may be a physical situation, which is perceived rather than described, and the conclusion either visualised or arrived at via physical experimentation. For example, given the "premises" *bananas hanging from the ceiling, planks of wood and bricks stacked in the corner*, and the goal "retrieve the bananas", the "conclusion" *planks across stacked bricks underneath bananas* may come in a flash of insight, or may require some messing around.[45]

If there is no precise starting point to a goal-directed thinking process then it falls under what Johnson-Laird calls *creativity*. He suggests that "an act of creation yields a product that is novel (at least for the individual who created it) and that satisfies some existing criteria or constraints…" (1988, p.446). These constraints can be more or less complete, depending on the domain in which the creator is operating. For example, a jazz improviser has reasonably complete (if tacit) criteria within which a certain degree of freedom exists, but a composer producing a written score has more room to manoeuvre, since the creative constraints are (usually) less restrictive. Like reasoning problems, creative processes proceed in a non-deterministic fashion, but unlike reasoning, "there is no clear starting point in the problem space for an act of creation" (p.447). Thus, the creator faces the problem of choosing where to begin. Johnson-Laird identifies three possibilities: a *neo-Darwinian* procedure, in which "arbitrary combinations of or changes in existing elements" are generated, and then put through the filter of the domain criteria; a *neo-Lamarckian* procedure, in which the "initial combinations of or changes in existing elements [are formed] under the immediate guidance of the criteria of the domain"; and a *multistage* procedure, in which one set of criteria is applied in the initial generative stage, and another set is applied at the filtering stage (ibid.).

This completes Johnson-Laird's taxonomy of thinking, which encompasses association, calculation, induction and deduction (understood broadly), and creativity. Naturally these kinds of thought are not mutually exclusive, indeed in the ordinary course of events they tend to intermingle quite freely, and are somewhat inter-dependent (for example, association has a role in creativity). Nevertheless, each of them, considered on its own, exhibits the features I've suggested are most broadly characteristic of higher-thought: sequences of conscious states; and a large degree of independence from the environment.[46] Notice also that, since these kinds have been specified at the task level, no commitment is being made here to any particular form of internal processing (beyond the broad features I've mentioned). What Johnson-Laird's taxonomy provides is, rather, a way of characterising

---

[44] Although, confusingly, he calls the process "deduction".

[45] A difficulty here is that it's not clear this kind of problem solving strictly admits of a distinction between induction and deduction, although it certainly does have a precise starting point, and proceeds in a non-deterministic fashion towards its goal. However, the gestalt tradition may provide a suitable replacement, viz: the distinction between *productive* and *reproductive* thinking, which can be loosely mapped onto induction and deduction, respectively. The former involves the generation of something new or novel, while the latter borrows from past experience, and/or recombines existing elements.

[46] The problem case is association, which may easily get caught in the grip of external stimuli, and hence fall outside my criteria for higher-thought. However, insofar as association forms part of another thought process (e.g., creativity or induction) then it is at least a *component* of higher-thought. Moreover, there are times when association does seem to proceed in a fairly autonomous fashion (think of lying in bed at night after drinking too much coffee).

in some detail the *targets* of cognitive science.[47] But this, in effect, is to ostend the higher thought processes themselves: they are the causal basis of the input-output profiles I've delineated, i.e., that sub-class of inner operations giving rise to the behaviours of interest. Notice also that, since the inputs and outputs are specified in terms of objects in the environment (e.g., strings of linguistic tokens, physical situations, and so on) and behaviours exhibited by the organism (e.g., utterances, actions etc.) – at what we might call the *organismic* level – then the operations of interest are likely to be *global*, *personal level* processes. That is, they will involve faculties distributed across the whole brain, and (as already suggested) implicate a good deal of conscious phenomenology. Thus, while the processes we hope to discover will presumably have a decomposition into sub-personal, modular, unconscious processes, talk of 'higher-thought' properly refers only to this global level of causal explanation.

Before considering the relationship between consciousness and higher-thought, I want to consider a trio of further properties standardly assigned to human thought which are often held to make a connection between thought and natural language (or between thought and some kind of linguiform inner symbol system – a LOT) either necessary, or extremely likely. The properties I refer to are those of *abstractness*, *productivity*, and *systematicity*. While some forms of thought may lack these, they are at least partly definitive of higher-thought.

Generally speaking, to say that thought is *abstract* is to say: (1) that it is capable of exhibiting a high degree of *focus*, i.e., that it can *abstract away from* the many details of a situation and zero in on some particular feature or property; and (2) that it generates representations with *special kinds* of (non-perceptual) contents. Consider first our capacity to focus on very specific features of objects or events. For example, when one looks at a cup it is possible to specifically attend to, say, the circular shape of the cup's opening, despite the multitude of other features present in immediate perception. Circularity is a property that is shared by many particulars, and it is moreover, only one of a (very large) number of properties instantiated by the cup.[48] Thus, one's thought is highly focussed (on the *circularity* of the cup's opening), but also partakes of the general (since circularity is general). Or consider the thought: *triangles have three sides*. However such a thought is represented in us, it exhibits a degree of focus present in sentences, but not in other kinds of representations, such as maps, diagrams, images and so on. The latter are too rich and ambiguous in their style of representation; they both lack focus, representing far too much (a picture of a triangle will represent the triangle as having a particular colour, for example), and do not distinguish between a number of different thoughts (e.g., *triangles have three sides*, *this triangle has three sides*, and *this triangle is green*, could all be attached to the one pictorial representation of a triangle).

Section 3.1 introduced the idea that the elements of phenomenal experience (which, by assumption, are the contents of certain mental representations) exhibit varying degrees of abstractness. This territory is sometimes negotiated with the distinction between *sensory* and *non-sensory* kinds of experience (Lloyd 1996). According to Lloyd, sensory experiences, unlike non-sensory experiences, are *modality specific*; *basic* (meaning that they are not constituted by or dependent on other elements or experience); relatively *few in number*; and *compulsory* (pp.65-7). Of course, what is being marked here are the ends of a continuum.[49]

---

[47] This way of putting things is due to Andy Clark (1993a, p.4).

[48] Indeed, from most viewing angles, the opening of a cup looks elliptical, so the circularity in question is doubly abstract, in that it is not even strictly speaking a visual property of the cup. It is inferred of the cup.

[49] For this reason I generally choose to speak in terms of *degree of abstractness*, rather than adopt the sensory/non-sensory terminology. The latter, I fear, encourages the natural tendency to reduce complex continua to a simple dichotomy. Note that abstract elements of experience *can* be sensory (think of depth perception, or object

There are many subtle gradations along the dimensions Lloyd proposes. For example, the sensation of visual depth is clearly at the sensory end of the scale, whereas the spiral form experienced in Figure 3.1 (see above) is somewhat more abstract, because it is *less basic* – information concerning the parts of which it is composed must be processed before a spiral gestalt can be generated. And the recognition that, for instance, a spiral has something in common with a circle, presumably relies on processes that occur still further from the sensory periphery. Moreover, while depth perception is normally a *compulsory* feature of visual experience, there can be a degree of voluntary control associated with the experience of perspective and form, as I demonstrated with regard to the inverting stairs (Figure 3.2), and the vase/faces (Figure 3.3). This looseness of fit between the concrete source of experience, and its more abstract manifestations, becomes even more pronounced when considering the variety of associations that a visual form can engender (it might remind one of a place, a person, an event, a sound, and so on).

When it comes to higher-thought, it is clearly the non-sensory end of the scale that is most relevant. Not only are the contents of higher-thought decidedly *non-basic* (since they often rely on multiple layers of prior processing); and *non-compulsory* (in the sense that similar starting points in concrete experience give rise to an enormous variety of very *dissimilar* thoughts); but they are often independent of any specific sensory modality. Human thought can generate representations of things that are not physically present, or past and future events (a property known as *displacement*), and involves highly abstract notions like necessity, causation, and democracy, to name but a few. Thus, higher-thought produces representations with *non-perceptual* or *non-sensory* contents. It is the capacity to generate such representations which is at least partly responsible for the focus exhibited by higher-thought.

Next I will consider the productivity of thought. Thought is often alleged to be *productive* in the sense that there is in principle no limit to the number of distinct thoughts we can think, i.e., thought is unbounded. In practice, of course, we will never think more than a finite number of this (potentially) infinite set, but this is due to two fairly mundane facts about us: 1) we are mortal, so time will run out for the generation of the set; and 2) we have some resource limitations, such as a finite memory – thus even without the time constraint we would still be restricted to producing only a bounded set of thoughts, since memory limitations prevent us from thinking thoughts of more than a given length or complexity. Clearly then, the productivity of thought does not refer to genuine unboundedness in our behaviour (or our mental life); if productivity is construed this way then we are only so under the idealisation of unlimited time and resources. What it does refer to is the (assumed) existence of an unbounded capacity or competence which is not fully expressible due to performance limitations. That is, the productivity of thought is an *empirical hypothesis*, the acceptability of which depends on "whether you believe the inference from finite performance to finite capacity is justified, or whether you think that finite performance is typically a result of the interaction of an unbounded competence with resource constraints" (Fodor & Pylyshyn 1988, p.34).[50]

The idea of an unbounded competence originally arose in the context of considerations (largely due to Chomsky) regarding our ability to produce and comprehend language. The

---

gestalts), but as they become dependent on more and more stages of prior processing, and less subject to stimulus determination, the label 'non-sensory' becomes appropriate.

[50] Presumably the latter inference is most plausible in cases where the finite capacity in question is large. But this is precisely what motivates the move to an unbounded competence in the case of both language and thought – the readily accepted claim that within our lifetimes we produce/comprehend an *enormous* number of thoughts and sentences.

next move is an inference to the best explanation of this alleged competence, which Chomsky takes to deliver a combinatorial system of inner symbols. Fodor and Pylyshyn draw the same conclusion with respect to thought (1988, pp.33-5). It's important to be clear about the two steps here: the hypothesis of a productive competence; and its explanation in terms of a combinatorial symbol system. While the latter certainly explains the former (it shows us how to generate an indefinite number of well-formed formulae for the expression of thought), it may be that there other ways to explain productivity (see van Gelder, 1990). Moreover, a number of theorists have expressed reservations about the claim that thought (or language comprehension and production) actually is productive. Despite this, I take it to be *prima facie* plausible that *some* thought is productive (in the sense of relying on an underlying competence), as consideration of our arithmetic abilities attests. When it comes to doing sums, for example, the limitations we experience really do seem to be performance related. Given pencil and paper most people can do summations *much* better than without, suggesting that memory limitations are the dominant constraint on in-the-head summation (at which most of us perform poorly). The marks on the paper are acting as an external memory device, which allow us to reveal more of our underlying arithmetic competence. Note that this is not intended as an argument for an inner combinatorial system of representation, but only for the existence of a particular generative capacity (see above).

Lastly I turn to the systematicity of thought. Thought is allegedly *systematic* in the sense that our capacity to think one thought is *intrinsically related* to our capacity to think other, structurally related thoughts. To take the standard example, one doesn't find people who can think that John loves the girl, but can't also think that the girl loves John. Another example: one can't think that John is likeable, and Bill is sympathetic without being capable of thinking that Bill is likeable, and John is sympathetic. The idea is that no-one has a *punctate* intellectual competence; potential thoughts do not occur in isolation, but come in structurally related groups. As Clark puts it: "our potential thoughts form a kind of closed set" (1993a, p.147). Like productivity, systematicity has been used to argue for a system of mental symbols with constituent structure (Fodor and Pylyshyn 1988, pp.37-41).

Systematicity does seem to be a property of linguistically-mediated thought. Unlike productivity it is something one can more or less check on by inspecting the thoughts we have (since it doesn't involve idealisation to an unbounded competence). But there are those who raise doubts about the systematicity of non-linguistic thought (i.e., thought in which representations of natural language tokens play no part). Is it really the case, for example, that the thought of infraverbal organisms is systematic through and through? Fodor and McLaughlin think so: "you don't find organisms that can learn to prefer the green triangle to the red square but can't learn to prefer the red triangle to the green square" (1990, p.184). But suppose a chimp can think that leopards are dangerous, and that bananas are delicious – do we really want to say that a chimp must also be capable of thinking that bananas are dangerous and that leopards are delicious?[51] Sterelny puts it this way: "We (perhaps via our public language) have achieved a separation of form and content, so some of our mechanisms for constructing thoughts are independent of domain, independent of what they are about. We have no grounds for assuming other creatures have achieved [this]" (1990, p.183). Since this dispute largely comes down to opposing intuitions (at this stage) it is reasonable to withhold judgement on the question as to whether non-linguistic thought really is completely systematic.

It's time to summarise what I've established regarding the nature of higher-thought. Recall that I started with a generic conception of thought common to all cognitive scientists: thought as semantically coherent causal processes defined over inner representations.

---

[51] This example is adapted from Sterelny 1990, p.183.

Higher-thought is a subset of these exhibiting a number of special properties. At the most general level it consists of processes that are largely internally-driven (i.e., not stimulus-locked) and that contain extended sequences of conscious states. More specifically, it consists of the processes responsible for association, calculation, induction and deduction, and creativity, as specified at the task level in Johnson-Laird's taxonomy of thinking. Such processes are likely to be global, and will require specification at the personal level (even if they are decomposable into sub-personal, unconscious processes). In addition, they will be abstract, and may exhibit the further properties of productivity and systematicity. These last two don't have the status of being necessary characteristics of higher-thought, since the existing evidence is equivocal in this regard. It may be that only those thoughts directly involving some form of natural language will have these properties.

Understanding Experience

Given this characterisation of higher-thought I'm in a position to reconsider the relationship between consciousness and thought. What has emerged as an important characteristic of higher-thought is that it generates *sequences* of conscious states. This suggests, contra Lashley, that our thought processes are transparent to at least this degree: that many of the intermediate steps they contain contribute to phenomenal consciousness (in contrast with the interlevels of representation generated by basic perceptual processes – assuming the standard account of these).

But it's not just the existence of sequences of conscious states that is most characteristic of higher-thought. Higher-thought also generates a phenomenology that is *qualitatively different* from the experiences connected with simpler forms of thought. Following Strawson (1994, pp.5-13) I'll call this phenomenology *understanding experience*. This species of experience belongs to the family of abstract experiences, in particular, it consists of all those varieties of consciousness that accompany higher-thought. Understanding experience is perhaps best exemplified in the phenomenology of linguistic thought. It is by no means clear that natural language is a requirement of all higher-thought, but it is equally clear that a good deal of higher-thought depends, in a non-trivial way, on the use of symbol systems of one kind or another. Linguistic thought may be among the highest expressions of our cognitive potential, and the most common source of understanding experience. I therefore propose to restrict my focus to linguistic thought from now on. There are some who will regard understanding experience in a dubious light, so I will also attempt, in what follows, to marshal some support for its existence.

Let me begin with speech perception. The sounds we use to communicate appear to be subject to a whole series of processing stages prior to the emergence of their meanings. The sonic stream must be segmented into phonemes, then morphemes (the smallest units of meaning), words, phrases and sentences. While it is not clear that these processes are completely sequential (because higher-level information can affect the interpretation of ambiguous or distorted lower-level information), it is nevertheless possible to discern a hierarchy proceeding from the more concrete to the more abstract aspects of speech. This process normally culminates in an understanding experience of some kind. Consider the difference between Jacques (a monoglot Frenchman) and Jack (a monoglot Englishman) as they listen to the news in French.[52] While there is a sense in which Jacques and Jack have the same aural experience, their experiences are utterly different in another respect. Jacques *understands* what he hears, while Jack does not. This difference is not just a difference in Jacques' capacity to respond to what he hears, it is a difference *within* phenomenal experience. Jacques consciously experiences something that Jack does not. It is this

[52] This example is taken from Strawson 1994, pp.5-6.

understanding experience that's missing when no sense is conveyed by what one sees or hears.

It's important to note that Jack may also fail to experience the French phonemic and morphemic boundaries that Jacques is privy to. Foreign speech is usually discerned as a relatively seamless stream of sound, in which individual words are difficult to pick out. This phenomenology is faithful to the sonic reality, since analysis of the sound waves associated with speech shows there to be no natural gaps or silences between words (Pinker 1994, pp.159-60). But even if one is familiar enough with a foreign tongue to succeed at word segmentation, there can still be something lacking in conscious experience. Think of an English speaker with high-school French, listening to complex French discourse (or for that matter a lay-person listening to two physicists discussing their work). Or, to borrow another example from Strawson, consider a code that consists entirely of English words which are used to stand for other words of English. Anyone familiar with the code will react to coded messages with an "automatic and involuntary understanding-experience" (1994, footnote p.6). The uninitiated, even though they share the basic auditory experiences up to, and including, word segmentation, will lack this kind of experience.

Reading text in a familiar language generates understanding-experience too. The existence of this element of consciousness can again be marked by highlighting cases where it is absent. A nonsense phrase, such as 'ilfaren noc eltranen', when read to oneself, certainly gives rise to the familiar experience of *sotto voce* inner speech, with clearly marked phenomenal boundaries between phonemes and words, yet it fails to generate any phenomenology of meaning. Or consider 'all mimsy were the borogoves'[53]. In this case the English words 'all', 'were' and 'the', create a grammatical (and experiential) structure in which *some* meaning is embedded (we know that all of the "borogoves" were "mimsy", whatever *that* means). But something is clearly missing by comparison with 'all flimsy were the borrowed hoes', which conveys something far more definite. One can even lack understanding-experience when exposed to strings composed entirely of bona fide English words. Try the following:

Fortunately honesty shook the boy obtained a big fault of them all things go away. …1)

This sentence is not grammatical English, and as a whole makes no sense, even though it gives rise to occasional hints or traces of meaning. Despite one's recognition of its component words there is no understanding-experience. Now consider:

Colourless green ideas sleep furiously. …2)

Odourless green frogs sleep quietly. …3)

The first, Chomsky's famous sentence, is grammatical, but conveys nothing. The second, with the same grammatical form, and the same feeling of grammatical "rightness", generates an additional element in consciousness – an experience of understanding or meaning.

There are several points to emphasise in all of this. First, the various contrasting pairs above lend considerable weight to the view that understanding experience is a genuine element of phenomenal experience. When one compares a nonsense phrase like 'ilfaren noc eltranen' with a good English sentence like 'men of few words are the best men'[54], it is hard to deny that the latter engenders a kind of experience the former does not. The difference is not as marked as that between, say, visual and auditory experiences, but it is real for all that. To deny this – to regard these phrases as distinctive merely because of their differential

---

[53] From Lewis Carroll's "Jabberwocky" in *Through the Looking-Glass*.

[54] From *King Henry V*, Act III, Scene ii.

effects on behavioural and cognitive dispositions – is not at all plausible. There is surely "something it is like" to understand a sentence of English.

Second, understanding experience appears to be composed of a number of distinct phenomenal elements. William James argues that we "ought to say a feeling of *and*, a feeling of *if*, a feeling of *but*, and a feeling of *by*, quite as readily as we say a feeling of *blue* or a feeling of *cold*." (1892). He further claims that:

> There is not a conjunction or a preposition, and hardly an adverbial phrase, syntactic form, or inflection of voice, in human speech, that does not express some shading or other of relation which we at some moment actually feel to exist between the larger objects of our thought. If we speak objectively, it is the real relations that appear revealed; if we speak subjectively, it is the stream of consciousness that matches each of them by an inward coloring of its own. (ibid.)

These feelings of relation are among the phenomenal elements that typically make up an understanding experience, which include: the experience associated with grammatical structure; the feeling of grammatical "rightness"; the phenomenology of individual words; and the meaning "gestalts" associated with whole phrases and sentences. Such elements are revealed in the kind of contrastive analysis I conducted above. For example, sentence 1) has some phenomenology of word and phrase, but fails to generate the experience of a total meaning structure that coheres with these parts. Sentence 2) feels grammatically right, but, unlike sentence 3), generates no meaning gestalt. This is presumably due to the strange juxtaposition of concepts it contains.

Finally, given its composite nature, it is no surprise to find that the varieties of understanding experience lie along some kind of continuum. We move by subtle degrees from a phrase like 'ilfaren noc eltranen' which displays the naked form of words not clothed with meaning; through 'all mimsy were the borogoves', which generates an unfilled meaning "gestalt", and 'shuffling elephant slowly hence', with rich traces of sense that aren't securely locked in place by syntax; to 'colourless green ideas sleep furiously', which has both a feeling of grammatical "rightness" and lots of free-floating phenomenology (attached to words that don't quite hang together); and finally a sentence like 'from there to here, from here to there, funny things are everywhere'[55], in which all these various components of understanding experience come together. Of course, this motley collection doesn't do justice to the full range and depth of possible understanding experiences. In ordinary discourse the use of metaphor adds a further dimension to our experience of language, and it's hard to convey the subtle phenomenology that good prose or poetry can beget. It's also important to note that these experiences, since they result from the interaction of linguistic input with the individual mind, are often quite idiosyncratic. None of us ever shares quite the same experience of a sentence.

Understanding experience is, then, a species of abstract experience, one that is particularly associated with linguistically mediated thought, and which, having a variety of component parts, comes in a series of "strengths". I note also that it is quite fleeting, and can be hard to identify. James classifies what I call 'understanding experience' among the "transitive" states of mind (which he contrasts with the "substantive" parts of the stream of consciousness). These states are difficult to pin down: "the attempt at introspective analysis in these cases is in fact like seizing a spinning top, or trying to turn up the gas quickly enough to see how the darkness looks" (1892). But it is a mistake to infer from this difficulty to the non-existence of transitive kinds of consciousness. As James says:

---

[55] From *One fish, two fish, red fish, blue fish* (Dr. Seuss 1960).

> If to hold fast and observe the transitive parts of thought's stream be so hard, then the great blunder to which all schools are liable must be the failure to register them, and the undue emphasizing of the more substantive parts of the stream. (ibid.)

3.3 The Nature of Phenomenal Experience

In this chapter a number of important features of consciousness have emerged. We've seen that phenomenal experiences arise both as a result of the rapid, mandatory processing that occurs within input systems, and in the service of the slower, more flexible processes of higher-thought. In the case of the latter, experience takes on a sequential aspect, since many of the intermediate stages of higher-thought have conscious contents. Of course the sequences of conscious states generated by higher-thought processes need not be temporally contiguous. The stream of thought, as William James calls it, is nothing if not diachronically plural, i.e., any one train of thought is likely to be continually slowed or halted by competing, intercalated processes.

We've also established that instantaneous consciousness is a complex aggregate of phenomenal elements, since both higher and lower thought processes are co-present in the moment by moment activity of the brain. At any instant we thus find ourselves the subject of visual, auditory, proprioceptive and tactile sensations, to name a few, in addition to a range of understanding experiences connected with our voluntary deliberations. Moreover, each of these modalities has a complex internal structure, comprising a range of more or less abstract phenomenal parts, from the relatively concrete experiences of line and tone in vision, through object gestalts and feelings of familiarity, right up to the highly abstract deliverances of linguistic thought. Understanding experience itself is structured, i.e., it is an aggregate of distinct parts, since we can distinguish among feelings of *and* and *if*, word and phrase gestalts, feelings of grammatical rightness, and so on, which combine to create a total sense of meaning. We also recognise in understanding experience the range of degrees of abstractness found in experience generally, since understanding can be relatively basic and compulsory (as when someone yells "Look out!"), or quite rich, complex and idiosyncratic (think of the way you respond to good poetry or prose).

In addition, it has emerged that ongoing experience is best conceived as a mass of parallel strands or streams. To speak of 'parallelism' here is to suggest that there is a degree of independence among the parts of experience. This is most strikingly evident *across* modalities, which we know can exist more or less independently of each other, and is also evident to some extent *within* modalities, as evidenced by the various deficit studies I've discussed. The nature of abstract experience suggests that certain kinds of dependencies do obtain among the more concrete and the more abstract parts of perceptual experience, namely: that the less basic elements (e.g., facial familiarity) are contingent on the successful processing of more basic percepts (such as line and form). But as we ascend to very abstract kinds of experience there is a loosening of the hold that concrete experience has over higher experience. Thus, there is a good deal of endogenous control over the contents of understanding experience – they are not stimulus-driven, but are the result of a complex interplay between external stimuli (which directly translate into basic experiences), and internal dispositions and goals.

I conclude that phenomenal consciousness is a multi-modal aggregate; a composition of distinct phenomenal elements. It is highly parallel, both inter-modally and intra-modally, and includes a great deal of abstract phenomenology, particularly in the service of higher thought processes.

4

The Unity of Consciousness

Mainstream accounts of phenomenal experience typically hold it to be both *unified* and *serial*. For example, Churchland tells us that "consciousness harbours the contents of the several basic sensory modalities within a single unified experience" (1995, p.214, italics in the original); Penrose says that "a characteristic feature of conscious thought…is its 'oneness' – as opposed to a great many independent activities going on at once" (1989, pp.398-399); and even Dennett, who develops a multiple drafts theory of phenomenal consciousness, eventually accepts that "conscious human minds are more-or-less serial virtual machines implemented – inefficiently – on the parallel hardware that evolution has provided for us" (1991b, p.218). From this perspective the account of consciousness presented in the last chapter is somewhat unorthodox, since I claim that conscious experience is both highly parallel, and in some sense *dis*unified (an aggregate of relatively independent phenomenal elements). In this chapter I intend to show that, to the extent that seriality and unity have a coherent and useful role in characterising consciousness, my account is not in conflict with them. But I'll argue that there is much that is incoherent, or plain false, in the way unity and seriality are standardly treated. Tracing the source of these mistakes will reward us with an insight into the range of choices open to theorists with a computational perspective on consciousness.

Consider the following view of consciousness, as expressed by Baars, in his influential book *A Cognitive Theory of Consciousness*:

> There is much evidence for the seriality of conscious contents, but it is difficult to prove that the seriality is absolute. Conscious experience is *one thing after another*, a "stream of consciousness", as William James called it. Psychological theories that are largely confined to conscious processes…postulate largely serial mechanisms. And, as Wundt observed in the 1880s, even two simultaneous conscious events are experienced either fused into a single experience or serially, one after the other. There is no such thing as true psychological simultaneity of two distinct events… (1988, p.83, emphasis added)

Apart from being a nice statement of a very common view of consciousness, this quote illustrates the way seriality and unity go very much hand in hand. To be serial conscious experience must be "one thing after another", the opposition here being between the one and the many. If many things in a system co-occur – if there is "true psychological simultaneity" – then that system is operating in a parallel, not a serial fashion. So the diachronic seriality of consciousness implies its *oneness* at each instant, that is, its unity.

Baars' understanding of consciousness is very widely shared. Nonetheless, there is a tension lurking here, which is evident in most discussions of the "unity" of consciousness. On the one hand, Baars tells us "even two simultaneous conscious events are experienced either fused into a single experience or serially, one after the other", but on the other he says that "there is no such thing as true psychological simultaneity of two distinct events". There is an obvious inconsistency here, as the first statement appears to be committed to what the second denies, i.e., the simultaneity of distinct conscious events. A way of resolving this inconsistency is to read Baars' talk of "two simultaneous conscious events" being "fused into a single experience" as a claim about simultaneous *contents*; namely, that different contentful

elements can be fused into a unified experience, and hence there is no "true" simultaneity of "distinct [conscious] events". Thus, while the experience remains "single", its contents are multiple. Unfortunately, the consistency we purchase in this way is bought at a high price, for we are left with no clear conception of the kind of unity on offer here. It can't be a monolithic unity, because the "single" experience Baars refers to is composed of two distinct contents. A consciousness of this sort is in fact some kind of composite, and its unity must consist in something other than true, seamless oneness. The only alternative seems to be to deny the possibility of fusion and assert that distinct conscious contents must always be experienced single file.

These two ways of conceiving conscious experience – as a somehow unified composite of distinct contents (I'll call this fusion), and as a serial stream containing only one conscious content at a time (I'll call this genuine seriality) – can be likened to varieties of unaccompanied (a cappella) choral music. Genuine seriality is like a solo performance, in which the chorus remains silent, and a single voice is all that we hear. The limitations of the human vocal folds ensure that such a solo is monophonic, i.e., it contains only one note at a time. An advocate of genuine seriality supposes that the brain imposes a similar limitation on phenomenal experience – it can contain only one informational content at each moment. A nice feature of this analogy concerns the duet. Operatic music often involves the musical equivalent of a dialogue, in which two singers alternately take the melodic line. Such music changes in tonal quality as each singer (say a soprano and an alto) takes a turn, yet it is monophonic throughout. Genuine seriality implies that we are only ever privy to a single mode of experience at a time (since we are only privy to a single content at a time), but, like the duet, consciousness can involve switching between modes, thus changing its "tonal" quality.

On the other hand, if we suppose that conscious experience incorporates a number of informational contents (perhaps from distinct modalities) fused into a unity, then it is best likened to polyphonic choral music. Polyphony involves two or more simultaneously active voices, such that at any moment there are a number of different notes being sounded. An individual voice is like a single mode of experience. Each contributes to the total sound, yet there is a sense in which none is entirely independent of the others, since they combine to form a balanced harmonic whole. In a similar fashion, an advocate of fusion supposes that the brain binds together a collection of simultaneously active, but distinct, informational contents into a unified consciousness. Exactly how this binding operation is achieved remains to be specified. One obvious suggestion is that unity is a consequence of the neural basis of consciousness being in some way singular (e.g., it might involve a single representational vehicle with a complex content, or a single consciousness-making process).

Both the monophonic and polyphonic models of consciousness are at large in the literature. But it is my view that neither is completely satisfactory, at least as they are typically understood. In order to defend this claim, I will consider them in turn.

4.1 Monophonic Consciousness

Penrose appears to be an adherent of the monophonic model. Consider the following:

> Utterances like 'How can you expect me to think of more than one thing at a time?' are commonplace. Is it possible *at all* to keep separate things going on in one's consciousness simultaneously? Perhaps one *can* keep a few things going on at once, but this seems to be more like continual flitting backwards and forwards between the various topics than actually thinking about them simultaneously, consciously, and independently. If one were to think consciously about two things quite independently it would be more like having two *separate*

> *consciousnesses*…while what seems to be experienced…is a *single* consciousness which may be vaguely aware of a number of things, but which is concentrated at any one time on only one particular thing. (1989, p.399)

Thus, according to Penrose, consciousness is "single", because we can't hold two independent thoughts at the same time. Doing so would be like having two consciousnesses in the one head. And while we sometimes appear to be managing several thoughts at once, this is really just timesharing, like Papageno and Papagena in their famous duet.[56]

There is some truth in these claims. However, read as a description of *consciousness* this material is mistaken on at least two counts. On the one hand there is evidence of a straightforward conflation of consciousness and what I earlier styled "higher-thought" (see Chapter 3). And on the other, there is an apparent confusion between consciousness and attention.

Consider first the nature of human thought. During our lifetimes we humans acquire some tremendously powerful cultural products, including a variety of languages and symbol systems. These enable us to conduct extremely abstract trains of thought, which, so far as we know, no other creature on the planet can reproduce. Such thought processes generate a good deal of conscious phenomenology, and not just the visual or auditory phenomenology associated with perceiving symbols, but also the understanding experience required for their proper manipulation.

So Penrose is right to identify an important relationship between consciousness and higher-thought. Moreover, there is something to be said for the claim that higher-thought is serial; we do seem, in some sense, to be restricted to a single "topic" at any given moment. However, it is surely not consciousness *in toto* that is so restricted. Even the most casual inspection of our instantaneous phenomenal field reveals that contents drawn from different modalities can simultaneously co-exist in experience. This is a point that really should be banal, but is often overlooked in discussions of consciousness. When you go for a walk in the country, for example, you not only have a great deal of pleasant visual experience, but at the same time you hear the sound of wind in the trees and birds singing, you feel your feet hitting the ground, and you have a sense of your bodily state (whether you're tired, energetic etc.). It is *not* the case that you, for example, first hear a bird, and then see it (or vice versa). Sound and vision don't compete for a place in awareness; both are simultaneously present to us.[57]

---

[56] *The Magic Flute,* Act 2.

[57] A possible reply, at this point, is that, while it might seem that we have more than one modality in consciousness at a time, this is an illusion generated by very rapid mode-swapping. It is hard to take this reply seriously. Apart from being rather ad hoc (why would one assert this, unless committed to genuine seriality, come what may?), it seems to conflate consciousness and attention (see the remarks to follow).

Baars sometimes gives the impression that he favours this view. For instance, he tells us that when sensory inputs arrive in the cortex they must "compete for access to [a] limited-capacity system". If an auditory and a visual stimulus arrive simultaneously we get "stimulus competition":

> One may be a speech sound in the left ear, and the other a falling glass in the right visual field…In our scenario, only one of the two can be broadcast at any moment, because they conflict in spatial location and content, so that the two simultaneous cortical events cannot be fused into a single, consistent conscious event. (1988, p.126)

This suggests that information from two modalities cannot "fuse" in consciousness – experience can only occur in monophonic mode. But it just isn't plausible to assert that "a speech sound in the left ear" and "a falling glass in the right visual field" are somehow inconsistent with each other. Right now I am aware of a computer monitor in front of me, and of sounds coming from behind me. I may not be able to simultaneously react to, or focus on, both of these sources, but they are both part of the present conscious moment.

Once one distinguishes phenomenal consciousness *per se*, and higher-thought, with its various phenomenal concomitants, it becomes clear that conscious thought is merely one component of a richer total experience. Not only does the phenomenology of higher-thought clearly fail to exhaust the possible contents of consciousness; it doesn't even *exclude* other kinds of conscious experience when it is in progress – while walking down the street I sometimes engage in quite complex conscious reasoning, but this phenomenology doesn't preclude me from *simultaneously* experiencing my surroundings in various ways. Thus, while it may be that I can only entertain one topic of higher-thought at a time, it is wrong to suggest that conscious experience as a whole is thereby rendered "single" or "serial". It is possible to admit the seriality of higher-thought, without accepting that phenomenal experience in general is monophonic. Penrose fails to spell out his views in such a way as to clearly distinguish the latter claim from the former.[58]

One way to make Penrose's position sound more plausible is to recast his discussion in terms of attention, that is, read him as suggesting that *attention* is serial (recall his claim that "a single consciousness" is "concentrated at any one time on only one particular thing" (1989, p.399)). For there does appear to be a mechanism whereby we can focus on some particular object or aspect of experience. And attention is clearly more restricted than consciousness in general. You are currently subject to a large range of phenomenal states, but your focus is on the task of understanding these sentences. If you shift focus to the sensation of the chair against your body, a process which heightens (but does not create *ab initio*) those very sensations, then your language understanding activities are temporarily, if briefly, suspended. Thus, attention does appear to be a process whereby first one "thing" and then another becomes focal.[59]

But one must be careful here. It is all too easy to slide from the claim that attention is serial to the claim that consciousness in general is monophonic. And this slide is by no means inevitable. A quite natural view of attention is that it concerns variations *inside* consciousness, implying that the content of focal attention is but one element of total consciousness.[60] Of course, if one denies this, that is, if one rejects the distinction between attention and total consciousness, then one does recover some kind of monophonic consciousness. A supporter of genuine seriality might make just this move, indeed, it is a forced move in the monophonic game. One can hardly have a varying focus *inside* a single-voiced consciousness, since the distinction between focus and periphery is, in part, a distinction between different, but contemporaneous, contents of consciousness. Attentional shifts within a serial stream are *without remainder*, i.e., they correspond to *total* shifts of consciousness from one object to another.

This conflation of consciousness and attention is very prominent in the literature, and is probably the chief motivation for adopting a monophonic model of consciousness. I propose

---

[58] There is some textual evidence to suggest that Penrose is aware of this distinction. He tells us that "oneness" is a characteristic feature of conscious *thought* (see above), and he has a tendency to refer to "thinking" rather than consciousness.

[59] This might explain Baars' claim that I cannot be simultaneously conscious of a speech sound in one place, and a falling glass in another; perhaps he's just pointing out that I can't attend to both of these at once.

[60] For example, when I play tennis my focus is on the movement of the ball, but I don't entirely cease to perceive other features of the environment (even if they become quite peripheral), and I'm still host to a complex mix of proprioceptive sensations, which enable me to maintain my posture and balance as I flit gracefully(!) about the court. Baars appears to support this kind of claim, asserting that when one focusses on, say, a piece of text, there is "much ancilliary information [that] is immediately available", some of which "is in the sensory periphery, like a kind of background noise" (1988, p.14).

to spend some time examining and criticising its sources, before turning to a consideration of the polyphonic model of consciousness.

Attention and the Monophonic Model

In his influential book *Cognitive Psychology* (1967), Neisser discusses a number of phenomena that have come to be associated with the monophonic approach to conscious experience. He claims that before the processes of focal attention can be brought to bear on particular figures in a visual scene, there are *preattentive* processes whose role is to segregate the figural units which later mechanisms will "flesh out and interpret" (p.89). According to Neisser, preattentive processes are involved in the involuntary control of attention, which is not directed at random, but is guided by cues that have already been extracted from the visual input, such as motion in an unattended part of the visual field. Preattentive processes are also involved in the guidance of movement. Neisser remarks:

> Most drivers have occasionally been startled to realize that they have not been paying attention to the road for the last half-hour. In walking, the same experience is so common as to arouse no interest. In these cases, the behavior has been steered entirely by the preattentive analyzers. These mechanisms are crude and global, and will not suffice for fine decisions; hence the driver must quickly become alert if a difficult situation arises. (ibid., p.92)

While Neisser tells us only that in this situation the driver has "not been paying attention", and that the preattentive mechanisms are "crude and global", many are tempted to add "unconscious" to the list. So called "unconscious driving" has been elevated to something of an empirical result (albeit a folk one) in discussions of consciousness. It is quite common to be told, when the powers of unconscious information processing are at issue, that "of course, one can *sometimes even drive unconsciously*!" The case of unconscious driving is a prime mover in the identification of attention with conscious experience.

A second kind of case which also has a role in this identification concerns our awareness of bodily states, and other "background" features. Baars, for example, tells us that

> In contrast to your conscious experiences, you are probably *not* conscious of the feel of your chair at this instant; nor of a certain background taste in your mouth, of that monotonous background noise, or the sound of music or talking in the background…(1988, p.3)

These kinds of examples are, if anything, even more prone to be regarded as the kinds of experiences that are only conscious *when attended to*, providing further (implicit) support for the claim that consciousness is coextensive with attention.

Usually this move, from such-and-such is *unattended* to such-and-such is *unconscious*, is made covertly, without argument or comment. However, occasionally such operations are conducted out in the open. For example, Mandler (1975), in his discussion of Neisser, is quite explicit about identifying consciousness with attention. Towards the end of *Cognitive Psychology* Neisser makes some fascinating remarks concerning the relationship between perception, memory and thought. He there develops the distinction between *primary* thought processes, which are "rich, chaotic, and inefficient", and *secondary* thought processes – those that are "deliberate, efficient, and obviously goal-directed" (1967, p.297). What Neisser appears to have in mind here is the distinction between the kind of thinking that goes on in dreams and fantasy, and deliberate, purposeful thinking (see p.298), which correspond, respectively, to *association* and *problem solving* in Johnson-Laird's taxonomy of thinking (see Chapter 3). Neisser goes on to suggest that "the processes of visual cognition, and perception in general, may serve as useful models for…thought". In particular, he claims that the primary process "constructs crudely formed "thoughts" or "ideas"," and so

functions like preattentive processes, while the secondary process of directed thought is like focal attention; it has the function of elaborating those objects generated by the primary process (pp. 301-304). Importantly, both primary thought processes and preattentive processes

> produce…fleeting and evanescent objects of consciousness, crudely defined and hard to remember. If their products are not seized on and elaborated by an executive process of some kind, they have little effect on further thinking and behaviour. (p. 301)

And later, concerning the primary process:

> [its] products are only fleetingly conscious, unless they undergo elaboration by secondary processes. (p.304)

Initially Mandler seems to go along with this characterisation:

> The products of the primary process alone…are only "fleetingly" conscious unless elaborated by secondary processes. By implication the elaboration by secondary processes is what produces fully conscious events. (1975, p.232)

But within the space of a few paragraphs he is claiming that the "processes that make up *consciousness* are secondary processes, secondary in elaboration and time to primary, preattentive processes that are *unconscious*…" (ibid., p.233, emphasis added). Indeed, he is so bold as to remark that we will "note why Neisser's contribution is important", if we "permit the free translation of 'attention' into consciousness" (ibid., p.232). He thus reconstructs Neisser's comment that walking and driving can be conducted "without the use of focal attention" (1967, p.92), as the claim that "[there] are processes that run off outside consciousness (unconsciously)" (1975, p.232).

Looked at purely as a matter of exegesis Mandler is clearly getting something wrong here. What Neisser tells us is that the products of preattentive perception and primary thought processes are "fleeting and evanescent objects of consciousness", which are "crudely defined and hard to remember". He *does not* say that they are unconscious, indeed he gives us a rough characterisation of the kind of conscious phenomenology these processes generate. Thus, even though driving can sometimes be performed "without the use of focal attention", this hardly amounts to the claim that it can occur "without the use of *consciousness*", at least so far as Neisser is concerned. Indeed, it is hard to imagine why anyone would want to claim this. But this brings us to matters of fact.

Block throws some light on the issue when he remarks:

> [when] the inattentive driver stops at a red light, there is presumably something it is like for him to see the red light – the red light no doubt looks red in the usual way, that is, it appears as brightly and vividly to him as red normally does. Because he is thinking about something else, however, he may not be using this information very much in his reasoning nor is he using this information to control his speech or action in any sophisticated way…(1995, p.241)

We might add – perhaps he is not storing this information in long-term memory. The claim here, which I fully endorse, is that it's surely not the case that an inattentive driver is completely unconscious of the activity of driving.[61] *At worst* the relevant objects of consciousness (the red of a traffic light, the shape of an approaching vehicle, the figure of a

[61] I take it no-one would claim that "unconscious driving" involves *complete* unconsciousness. A driver who isn't attending to his driving is presumably conscious of that which he *is* attending to, say, what he's going to cook for dinner when he gets home.

pedestrian) are "fleeting" and "crudely defined" – not entirely absent. More likely an inattentive driver actually focuses quite closely on his driving at times, if briefly, rapidly returning to his higher-thoughts when the task demands are reduced. That is, inattentive driving probably involves focal timesharing, such that the task of driving receives occasional slices of close attention. For the remainder of the time the objects of driving may be peripheral, but they are *not* unconscious.

So why is there the temptation to suggest otherwise? Block's proposals are helpful here. When one drives inattentively those conscious contents associated with the road appear not to be employed in reasoning, or in the sophisticated control of speech and action. This accounts for their sometimes fleeting character, and it also explains why they are so easy to dismiss. As language users, with the capacity for high levels of abstract thought, we find it tempting to identify as conscious only those experiences correlated with the highest levels of self-awareness, rationality and voluntary activity. Since inattentive driving involves many fleeting, ill-defined objects of consciousness, which evade ready description, and which are too ephemeral for use in conscious reasoning, they hardly seem like ours at all – they barely belong to us. No sooner have they served their purpose than they vanish without trace. Hence, it is easy to dismiss them entirely. The point is that the conscious contents generated while driving inattentively *don't have many effects*. More particularly, they don't hang around long enough to leave traces in long-term memory. And this remains the case whether they are vague and ill-defined, or vivid and intense. Thus, so called "unconscious driving" may simply be fully conscious, but rapidly forgotten activity; activity that leads to no higher reflections, or integrations, or any other form of inclusion in our more extended projects.[62]

Turning now to the various "background" phenomena cited by Baars, you will recall that he lists "the feel of your chair", "a certain background taste in your mouth", and "the sound of music or talking in the background", as being among the phenomena of which you are probably not conscious at this instant. Before questioning this claim, I note that Baars has failed to distinguish two significantly different kinds of cases. In Section 3.1 I drew attention to the fact that sensory neurons differ according to the speed with which they adapt to unvarying stimuli. Rapidly adapting (*phasic*) receptors, quickly lose their sensitivity to stimuli, and are thus suited to alerting us about changing conditions, while slowly adapting (*tonic*) receptors, are suited to providing information about steady states of the body. Gustatory (taste) and olfactory (smell) sensations, for example, involve phasic receptors.

---

[62] There is an important sense in which driving, walking, and other forms of motor activity genuinely implicate the unconscious. Everyone knows what it is like to consciously and laboriously learn the motor commands required to perform certain skilled behaviours. And everyone also knows what it is like to achieve a level of unconscious control of such behaviour. In fact, mastery requires unconscious control; it requires the motor commands to dip beneath the surface of consciousness. It is reasonable to suppose that the long periods of assiduous repetition involved in learning motor sequences result in the training of specialist systems to carry out those sequences. Think of learning to touch type, or to play a musical instrument, for example. Rather than taking individual letters/notes as inputs, the relevant specialists respond to *groups* of letters/notes (i.e., words/musical phrases). That is, as one's performance improves the whole process becomes more coarse-grained; "chunks" of input produce "chunks" of output, and one is no longer conscious of the individual muscle movements involved. (See Section 7.3 for further discussion.)

Mandler is therefore right to claim that "[there] are processes that run off outside consciousness"; the processes whereby we control complex, "chunked" behaviours are not available to us. But this isn't to suggest that skilled behaviours, performed without focal attention, are entirely lacking in phenomenology (which Mandler implies). The underlying process may be unconscious, but it appears to be directed by conscious states, and certainly generates a good deal of phenomenal experience. When driving, for example, it is the experience of an object in the periphery of vision that typically initiates a braking sequence. And the motor sequence itself has a characteristic phenomenal "feel". Such phenomenology varies considerably with the degree of driving mastery that has been attained, but for all that is undoubtedly an authentic element of consciousness. See Chapter 7 for more on this.

There is about a 50% decrease in the responsivity of olfactory receptors in the first second or so after stimulation, and similar figures apply to gustatory receptors. With continuous, unvarying stimulation, complete adaptation of taste sensations occurs in one to five minutes, meaning that no further experience is generated by the particular stimulant involved.[63] To speak of an "unconscious" experience is, presumably, to refer to the presence of sensory information that is capable of being rendered conscious, perhaps by a suitable shift of attention. But with respect to taste, it appears that there is no such information to recover (once a sufficient period has elapsed). Pressure sensations, on the other hand, are longer lasting, due to the involvement of tonic receptors in generating this kind of experience. Therefore, the claim that "you are probably *not* conscious of the feel of your chair at this instant" is open to being construed as a claim about unconscious contents. Specifically, it is interpretable as the assertion that your brain unconsciously represents the effect of the chair on various parts of your body, which information can be brought to consciousness at will. Similar remarks apply to sonic information.

But is this the right thing to say about sensations of pressure, or, for that matter, about unattended background sounds? Are they really unconscious? In the case of background sounds, for example, isn't it rather the case that when one first attends to, say, the sound of the cooling fan housed in one's desktop computer, one realises that this sound has been present in experience for some time? Because unattended, such a sound has not been labelled, or integrated into one's higher thought processes in any way (in conventional language: it has gone "unrecognised"), but it has nevertheless been an ongoing part of total consciousness. Or consider proprioception: the sensory feedback that emanates from one's limbs, creating a sense of their relative positions. Proprioception is even more prone to be relegated to the unconscious background, from which (it is supposed) it can only be retrieved by deliberately focussing on one's body. Yet, I would suggest, this sense of body position, like our externally oriented senses, is an ever present feature of experience.

Our consciousness is not monophonic or single. It is a complex amalgam of many elements, which, for the most part, are so constant that it's easy to take them for granted. We know of the persistence of visual experience, for instance, because we are all familiar with the decrement in phenomenology that accompanies closing our eyes. But most of us require a more striking demonstration than this in order to acknowledge the persistence of proprioception. Sadly, nature occasionally obliges in this regard. Sacks describes the tragic case of a woman who, due to acute polyneuritis of the spinal and cranial nerves throughout the neuraxis, suddenly loses her capacity to have proprioceptive experiences: "Something awful's happened," she tells Sacks, "I can't feel my body. I feel weird – disembodied." (1985, p.44). This woman has *none* of the usual (proprioceptive) feedback from her body. Without it she recognises (perhaps for the first time) what she had, but has now lost: the feeling of embodiment. Most of us don't realise that we *don't* feel disembodied, but she is in the horrible position of having this realisation forced upon her. The experience of embodiment, like a number of other "background" phenomena, is a constant feature of consciousness.

An advocate of the monophonic conception ignores, or misses, these constants of experience, and in so doing expedites the conflation of consciousness with attention. Once we reject that conflation, as I've argued we must, then the seriality of attention can no longer infect total consciousness. This still leaves us in need of some account of attention. For within consciousness, with all its richness, there are notable temporal variations – an

---

[63] The speed of sensory adaptation with respect to both taste and smell is too rapid to be accounted for purely in terms of the drop in responsitivity of the relevant sensors. While these initially undergo a rapid drop in sensitivity, the ensuing decline is more gradual, and a central mechanism of some kind is probably also required in order to account for the rapidity of the sensory adaptation.

enhancement here, a decrement there – and these must be explained. Neisser has proposed that we view attention as "an allotment of analyzing mechanisms to a limited region of the field" (1967, p.88). The idea is to treat attention as a matter of resource allocation; to suppose that there are simply not enough computational resources to permit a detailed inspection of, say, the whole visual field at once. Instead, preattentive processes provide a sketchy analysis, which attentive mechanisms then "flesh out and interpret" (ibid., p.89). That is, mechanisms of attention subject information already extracted from the world, and already displayed in the phenomenal field, to more intense processing. Such additional processing perhaps then generates the enhanced phenomenology that accompanies attention shifts. (Jackendoff has developed this kind of account – see his 1987, pp.280-283.)

An account like this is compatible with the claim that attention concerns variations *inside* experience. It also allows us to assign the seriality that some ascribe to the entire phenomenal field, to a *proper part* of consciousness: that element of experience associated with the current object of attention. Thus, we can accept that the focal centre of experience is, in some sense, univocal, without denying that it has a rich, polymodal periphery. One must be wary of misdescription even here. Penrose speaks of consciousness (read "attention") having a single "topic"; but how "single" is a topic? Attention shifts appear to be under both voluntary and involuntary control. As remarked earlier, movement in the periphery of vision tends to grab one's attention involuntarily, but one can equally direct attention where it is needed, in pursuit of some higher project. When one does so there are at least two kinds of changes to consciousness. First, there is enhanced detail with regard to the focal object, whether it be of sensory origin (e.g., the pressure in my left foot, the first letter of 'quaint'), or endogenously generated (e.g., a visual image, or a linguistic thought). And second, there is added phenomenology connected with the voluntary nature of the attentional act: higher-order thoughts and understanding experience, expressing a more abstract appreciation of the focal object (one might, for example, think "Oh, there's that pin I've been looking for!"). All this suggests that a "single" focal object may actually comprise a multiplicity of distinct elements. When one attends closely to a speaker, it's not just her words, but her meaning too, that become focal; when one listens to music, it's possible to attend not just to individual instruments, but to complexes of sound (the entire rhythm section, for instance).[64, 65]

---

[64] It doesn't strike me as unreasonable to suggest that in some circumstances we are even able to attend, simultaneously, to objects in distinct modalities (say, a sound and an image), even if we can't physically react to both (the sound might come from behind, and the image from in front).

[65] A suggestion regarding the confusion about seriality is that it really amounts to a poor choice of terminology. 'Serial', as a term of art within computer science, is understood by contrast with 'parallel'. A parallel process is one in which multiple computational instructions are completed each clock cycle, whereas a serial process can produce, at best, a single result per cycle. Most modern digital computers are restricted to serial processing, because of the bottleneck created by their CPU. True parallelism requires multiple CPUs (disregarding recent dual-issue chips, such as the Pentium). Thus, the atomic units around which the serial/parallel distinction revolves are the members of the von Neumann instruction set. Talk of attention being serial in virtue of taking one "topic" or "subject" at a time is nowhere near this precise. A single topic can be complex, with a number of concurrent phenomenal parts. Hence, a conscious thought process that takes one topic at a time, can also be regarded as parallel, in some sense. It all depends on the grain-size one adopts when drawing the distinction.

One way around this difficulty is to adopt a new term. When referring to attention, and to high-level conscious thought processes in particular, it might be more apt to use 'seriatim', rather than 'serial'. The former means 'taking one subject…after another in regular order' (Concise OED). Since a "subject" can be complex I think this captures what people really have in mind when they refer to thought as 'serial', namely: that conscious (high-level) thinking only takes one topic at a time, and proceeds in an orderly fashion. A seriatim stream of thought need not be serial, in the sense of comprising a unary sequence of phenomenal primitives; rather, at each moment the stream may contain distinct, parallel strands (sensory *and* cognitive), all contributing to focal awareness.

Serious doubt has been thrown on the claim that phenomenal experience is coextensive with attention. Neither inattentive driving, nor Baars' "background" phenomena really undermine the view that consciousness is a complex aggregate of phenomenal elements, both within its rich, polymodal periphery, and also, perhaps, at its focal centre. Together with my earlier remarks about the similarly mistaken conflation of consciousness with higher-thought, these considerations significantly diminish the appeal of the monophonic conception of consciousness. In the end, even if one isn't convinced by the analysis here, I would have thought that the simple case of sound and vision is a sufficient refutation of the monophonic model. The baby screaming in my left ear at this moment certainly diminishes my capacity to compose these words, but the sound he makes doesn't render me blind or incompetent! We are thus in need of an alternative to the monophonic conception of consciousness – which brings us to the polyphonic model.

4.2 Polyphonic Consciousness

It seems that the monophonic conception of consciousness is untenable. Many theorists recognise this, and so prefer to adopt a polyphonic model, in which many "voices" sound their notes simultaneously. Baars gives the impression that he allows for the "fusion" of a number of distinct conscious contents into a single, composite experience, but only so long as they are compatible kinds of contents (see above). Churchland advocates a more straightforward polyphonic model of consciousness. He includes in his enumeration of the "salient dimensions of human consciousness" the following:

> Consciousness harbors the contents of the several basic sensory modalities within a *single unified experience*. A conscious individual appears to have not several distinct consciousnesses, one for each of the external senses, but rather a single consciousness to which each of the external senses contributes a thoroughly integrated part. (1995, p.214)

This corresponds to what I have described as polyphony, because, while Churchland grants that *each* of the external senses contributes a part to consciousness, so rendering it "polymodal" in character (ibid. p.222), he nevertheless claims that these are found within a "single unified experience".[66]

In order to account for this, and a number of other features of consciousness (pp.213-24), Churchland develops the conjecture that phenomenal experience is the preserve of a particular neuroanatomical structure in the brain: the intralaminar nucleus in the thalamus. This structure has axonal projections to all areas of the cerebral hemispheres, and receives projections from those same areas. The brain thus contains a "grand informational loop" that "embraces all of the cerebral cortex", and which "has a bottleneck in the intralaminar nucleus" (ibid., 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". He does so as follows:

> There is one widespread recurrent system with an information bottleneck at the intralaminar nucleus. Information from all of the sensory cortical areas is fed into

[66] Note that a polyphonic theorist is not, strictly speaking, committed to a polymodal account of experience. A conscious event within a single modality can be complex, encompassing a number of distinct contents. Thus, one could advocate monomodal polyphony: the view that, while instantaneous consciousness embraces numerous distinct contents, these necessarily belong to a single modality. I don't think many theorists would be prepared to defend this position.

> the recurrent system, and it gets jointly and *collectively* represented in the coding vectors at the intralaminar nucleus, and in the axonal activity radiating outward from there. The representations in that recurrent system must therefore be *polymodal* in character. (ibid., p.222)

So information is conscious when it is represented by "coding vectors"[67] that lie within the information loop centring on the intralaminar nucleus. It is unified because these "polymodal" vectors occur within a single recurrent system.

This account is strikingly similar to Baars "Global Workspace" model of consciousness (1988). Baars' approach begins with the premise that the brain contains a multitude of distributed, unconscious specialist processors all operating in parallel, each highly specialised, and all competing for access to a global workspace – a kind of central information exchange for the interaction, coordination, and control of the specialists. Such coordination and control is partly a result of restrictions on access to the global workspace. At any one time only a limited number of specialists can broadcast global messages (via the workspace), since different messages may often be contradictory. Those contents are conscious which 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.)

In support of this global workspace model, Baars claims there is a brain structure suited to the role of workspace, namely: the Extended Reticular-Thalamic Activating System (ERTAS), which includes the reticular formation, the thalamus, and the "diffuse thalamic projection system" (p.124). The latter corresponds to the recurrent loop that Churchland takes to be so significant. The ERTAS is particularly suited to the role of "global broadcaster", given the bidirectional projection system that it incorporates. Moreover, there is "evidence of a feedback flow from cortical modules *to* the ERTAS" and of global information feeding "back into its own input sources". Baars suggests that "[both] kinds of feedback may serve to strengthen and stabilize a coalition of systems that work to keep a certain content on the global workspace". That is, given the competitive nature of access to consciousness, "a circulating flow of information may be necessary to keep some contents in consciousness" (p.134).

So both Churchland and Baars give informational feedback a pivotal role in their accounts of consciousness. Both identify brain structures that may act as a conduit – a functional bottleneck – through which information passes in order to become conscious (the thalamic projection system and associated structures), and both conjecture that these brain structures realise an executive system that guarantees the unity of consciousness. Their reasoning here seems to be as follows: given that consciousness, despite its polyphony, is just *one thing*, there must be *one thing* that underlies it; but it is implausible to suppose that the various distinct contents of instantaneous consciousness are encoded in a single representational vehicle; hence, there must be a *single* consciousness-making mechanism, an executive system that transforms the seeming plurality of polyphonic experience into a unity. The unity of consciousness, on this story, is not imposed by the *seriality* of the stream of contents broadcast (as would be the case with a monophonic executive), but by the fact

---

[67] Churchland seems to have in mind here both firing patterns within neural networks, and patterns of signals passing down axons.

that all the contents of consciousness pass through a *single* informational bottleneck and are subsequently broadcast throughout the brain.[68]

I argued in the previous section that our instantaneous phenomenal experience is typically polymodal in character, so the polyphonic accounts of Churchland and Baars are certainly an improvement over the monophonic model. However, their choice of an executive model of consciousness should give us pause, because this model is not at all plausible from an evolutionary perspective. Executive systems are dependent on a high degree of complex, system-wide integration. Everything we know about evolution suggests that nature tends to avoid such centralised, resource expensive solutions. While there are aspects of human consciousness that no doubt depend on the integration and co-ordination of information from multiple sites, there is little reason to think that consciousness in general is so constrained. Basic sensory experience is surely a product of simple, localised mechanisms. These theorists, because they take themselves to be committed to a single consciousness-making mechanism in order to explain the unity of consciousness, have overlooked a more parsimonious approach, one that is more consistent with our moment by moment phenomenal experience, and more natural from a neuroscientific perspective.

Polyphony: Single-track versus Multi-track

To see that an alternative account of consciousness is available, we need to explore the phenomenon of polyphony in a little more depth. Recall that choral polyphony involves two or more *simultaneously active* voices, such that at any moment there are a number of different notes being sounded. Each voice makes a distinct contribution to the total sound, yet none is entirely independent of the others. But how are we to understand this latter dependency? There are two possibilities. On the one hand we might think of it in terms of the superposition of sound waves, which results in a single complex wave reaching the listener's ear. On this reading, while diverse voices are involved in the creation of the work, the final product is a single physical structure. On the other hand, we might think of the dependency between the voices in terms of the harmonic whole formed by their co-occurrence, a harmony that can be broken by the alteration of any one voice. According to this quite different reading, the product is unified by virtue of the *connectedness* and *coherence* of its constituent parts, despite these parts having physically distinct sources.

Obviously, it is the first reading here that is more in tune with Churchland's and Baars' polyphonic models: instantaneous consciousness is unified in the sense of being a single broadcast. But there is still the second possibility. Suppose we treat the unity of consciousness not as a matter of physical oneness, but as a matter of harmony or coherence, a property manifest both in the consonance displayed by the representational contents of the various modalities (we see our bodily parts in positions we feel them; we hear sounds emanating from objects in the direction we see them; we taste the food that we can feel in our mouths; and so on), and in the binding of phenomenal elements within modalities. If we understand unity in this way there is no longer any need to deny the patently manifold nature of consciousness. Our phenomenal experience is a complex amalgam of distinct and separable conscious events; not a serial stream, but a mass of tributaries running in parallel. And it presumably has this form precisely because it is not the preserve of a single, executive computational device, but the combined output of myriad consciousness-making mechanisms distributed right across the cortex. This implies, pace Churchland, that a

[68] Another reason for adopting this strategy is that it enables one to account for the "limited capacity" of consciousness: the supposed fact that we are only conscious of a small fraction of the information our brains are generating at any given moment. I take up this issue in Chapter 5.

conscious individual does not have a "single consciousness", but a number of distinct phenomenal consciousnesses, at least one for each of the senses.

I can't resist employing another musical metaphor here to further distinguish this alternative polyphonic model of consciousness from its more orthodox counterpart. Before the advent of modern studios the only way to record music was to get all the musicians in a room together, place a microphone in their midst, and start up the band. The signal from the microphone would then go through a limited amount of processing before leaving a groove on a wax disc, or (more recently) a magnetic trace on a tape. Such a recording is said to be *single-track*, since there is no way to separate out the individual contributions of the musicians – they are packaged into a single structure. By contrast, on a *multi-track* recording one or more separate tracks are devoted to each musical instrument. As a result, while the sound produced at playback is a seamless combination of musical parts, *at the level of the recording* one can distinguish between them. For the sound engineer this has the benefit of allowing parts or voices to be independently brought in and out of the mix, so that their individual contributions to the total sound can be assessed.

The models of consciousness proposed by Churchland and Baars are single-track, because these theorists rely on a *single* consciousness-making system; a move which is forced on them by a literal reading of 'unity'. But if we discard this naive reading, and instead conceive unity in terms of coherence, then there is room in the theoretical landscape for a *multi-track* polyphonic model of consciousness. According to this model, each of the many discernible elements of instantaneous experience is, like the separate parts of a multi-track recording, the product of a localised, physically distinct structure or mechanism. Such a model is plausible not only from an evolutionary perspective, given that it avoids the need for a single, resource-expensive executive, but it also makes perfect sense of the available neurological and phenomenological evidence.

Consider, first, the phenomenological evidence. As I demonstrated in Chapter 3, close attention to consciousness reveals it to be a complex *aggregate* of many elements, each of which has a distinct existence. We know this, in part, because there are elements of experience that can be readily effaced without affecting the rest. When you close your eyes, for example, visual consciousness disappears, but the other elements of experience are unchanged: you can still feel where your limbs are, you can still hear the sounds coming in through the open window, and you can still feel the chair pressing against you body. Total deficits in sight and audition are quite common, and can be brought on suddenly by localised damage which leaves the other modalities more or less intact. This independence among the parts of experience is also evident *within* modalities, as we have seen. Recall the "looseness of fit" between the more abstract and the more concrete parts of visual experience. We have a degree of voluntary control over the appearance of the vase/faces illusion, for instance (see figure 3.2), but whether we see it as a pair of faces, or as a vase, there is no change in the underlying experience of line and tone itself. There is a common, primary visual experience, to which an additional variable element of abstract phenomenology is added (an object "gestalt"). Understanding experience also reveals itself as a distinct part of consciousness, because it is laid over a background of more concrete phenomenal elements, elements which can exist in the absence of the higher component (think of Jacques, the monoglot Frenchman, and Jack, the monoglot Englishman, listening to the news in French). Thus, we find that the parts of experience are like the parallel tracks on a multi-track recording – each track adds to the mix, but, since they run side by side, the loss of any one track does not affect the others, it merely reduces the total sound. Moreover, like a sound engineer, it seems we have some control over which parts get into the mix.

The real force of this phenomenological evidence for multi-track polyphony only fully emerges when it is conjoined with the available neuroscientific evidence. We know, on the

basis of deficit studies, that the information processing that supports conscious experience is realised in structures distributed right across the brain. And the distributed nature of this information processing is both an intra-modal and an inter-modal affair. Consider, again, our visual experience. Recent work in the neurosciences has shown that visual processing is highly modularised; the visual cortex appears to contain separate subsystems for the processing of information about colour, shape, depth and even motion (see Chapter 3 for more detailed discussion). 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. Recall the case (cited in Zeki 1993) of a woman who, as the result of a cortical lesion outside the primary visual area, lost the ability to detect motion visually. In spite of this deficit, her perception of other visual attributes appears to be normal. And recall, also, the strange phenomenology of prosopagnosia. Those who suffer with this condition, which arises in connection with very localised lesions, are unable to recognise familiar faces, despite having vision that is normal in all other respects, and despite retaining the capacity to *describe* faces.

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 – there is no need to advert to a re-presentation process elsewhere in the brain.

There are echoes here of Dennett's multiple drafts theory of consciousness (1991b, 1993). 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" (1991b, pp.253-4). 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.[69] Thus, Dennett, like me, advocates a multi-track model of consciousness.

To re-iterate: the plurality and neural distribution of cognitive processing modules revealed by deficit studies is best interpreted as evidence for the multiplicity of consciousness-making mechanisms in the brain. This is the hardware implementation of multi-track polyphonic conscious experience. Just as with the separate codings of the various voices on a multi-track recording, it accounts for the relative independence of the various strands of experience, and the "looseness of fit" between abstract and more concrete experiences within modalities.

---

[69] Given all this, it is initally surprising to find Dennett concluding that a conscious human mind is a "more-or-less serial virtual [machine] implemented – inefficiently – on the parallel hardware that evolution has provided for us" (1991b, p.218). This "von Neumannesque" virtual machine, he tells us, is "a product of cultural evolution that gets imparted to brains in early training"; "an organised and partly pretested set of habits of mind", whose major structural characteristic is "*serial chaining*, in which first one "thing" and then another "thing" takes place in (roughly) the same "place"." (pp.219-21.) However, Dennett's frequent use of the formulations "*human* consciousness" and "conscious *human* minds" (my emphasis) in this context, suggests that he takes von Neumannesque style consciousness to be merely *one component* of the total human experience, albeit a component that is probably peculiar, on this planet, to *homo sapiens*.

Coherence and Binding

Despite the attractiveness of a multi-track polyphonic approach to consciousness, there are a couple of obvious objections to this model. The first concerns the fact that multi-track polyphony seems to render coherence a *happenstance* of consciousness, rather than one of its defining features, because it places coherence at the mercy of a great many localised consciousness-making mechanisms. Unity (understood as coherence) begins to look like a contingent, rather than a necessary property of consciousness. The second is that multi-track polyphony seems to leave one vulnerable to the so called "binding problem": the problem of *explaining how* spatially and temporally distributed content-bearers can generate the manifest spatial and temporal coherence of phenomenal experience. However, a close examination of the issues raised here will not only demonstrate that the multi-track theorist has the resources to meet these objections, but will also serve to highlight the superiority of the multi-track approach over its single-track rival.

In order to convey the first objection more clearly let us again consider the nature of choral polyphony. A group of singers typically attempts to satisfy two kinds of constraints. On the one hand there are *temporal constraints*: choral voices must conform to a common measure of time, i.e., the notes each singer produces must start and finish in strict accordance with the beat. On the other hand, there are *harmonic constraints*: the singers must conform to a common scale so that the choral polyphony constitutes a pleasing noise at each moment, and follows a coherent path through the tonal landscape. However, it is somewhat misleading to describe polyphony in these terms. Strictly speaking, polyphony is just a matter of there being *multiple voices*. Neither harmony, nor adherence to the beat, are *necessary* features of choral music; they can be accidentally lost, or purposely discarded in pursuit of avant-garde tastes. Indeed, given that each singer constitutes a separate sound-source, musical unity is at the mercy of *each and every voice*. Both temporal and harmonic control is distributed among the singers, and musical coherence is a matter of the co-operative exercise of individual capacities (subject to the ready availability of information concerning the current state of play).

To increase the reliability of choral performances, one can centralise control of the singers to some extent. This, of course, is the role of a conductor: instead of relying on a bunch of people to independently keep time, everyone agrees to take their cue from a single time-keeper. But if it's a cast-iron guarantee of temporal and harmonic coherence one is after, then the only real option is to do away with the choir. A first option is to swap choral for instrumental polyphony, say, hire a pianist. In this context, a piano effectively constitutes a single sound source, and one that is subject to central control. More radically: one could use a modern polyphonic synthesiser (which is capable of producing a wide range of sounds, including realistic vocal tones), linked to an on-line sequencer/composer program. The output of a synthesiser is a single, complex, oscillating voltage, which is then transduced (via amplifier and speaker) into sound waves. By interposing sequencer between player and output, it is possible to smooth out timing errors (if one is prepared to live with a small delay), and even correct for poor musical judgement. Thus, prior to the generation of any sound, the musical input (the player's contribution) is filtered and modified by a central composer, so as to conform to some pre-determined standard of musical coherence. Although the sonic product is still recognisably polyphonic, the multiple, independent sound-sources of the choir have been replaced by a single waveform, the shape of which is subject to strict control by a master decision-maker.

This is the kind of approach that single-track theorists take towards consciousness, because they posit a global consciousness-making system (a composer and broadcaster) whose role is analogous to that of the electronic sequencer. Single-track models seem to be at

least partly geared towards guaranteeing the coherence of phenomenal experience.[70] Multi-track models, on the other hand, cannot underwrite such a guarantee, given that they presuppose a multitude of physically separate and relatively independent consciousness-making mechanisms distributed about the brain. Local failures are capable of undermining coherence in a way that is not possible on a single-track model. It is in this respect that multi-track polyphonic consciousness is best understood by analogy with the polyphony of a choir.[71]

So supporters of a multi-track polyphonic model of phenomenal experience seem to be committed to the view that *coherence* (both inter-modal and intra-modal) is a contingent feature of consciousness; a feature that is capable of lapsing *without thereby extinguishing consciousness*. Obviously, if such lapses were never observed to occur, the multi-track model would be in a difficult position. But, in fact, the existing evidence, far from undermining the multi-track model, tends to make the single-track model look misguided, at least insofar as it is adopted in order to guarantee the coherence of consciousness. I begin with the inter-modal case.

Consider the experience of a synesthete. *Synesthesia* is the involuntary stimulation of sensation in one modality by input to another. Most commonly it takes the form of "hearing colours", in which colour sensations are evoked by sounds. Those who are subject to this strange form of life, report a consistent and reproducible pattern of colour/sound associations, although there are few similarities across subjects. Michael, a synesthete whose case is documented in *The Man Who Tasted Shapes* (Cytowic 1993), has a more unusual form of synesthesia in which tactile sensations are generated in response to tastes. These sensations are experienced as notional objects, close to the body, which Michael can literally reach out and touch. The following is reported by Cytowic immediately after giving Michael a shot of Angostura bitters (conducted as a blind test).

> He shivered past the bitter part and spoke quickly. "Yes, the round part comes first, with a spongy texture," he said, tracing a curve with both hands this time. "Then the shape develops – I feel the holes now," he said, closing his fingers. "Here are the strands. A little thread. It gets bigger, like a rope. If I pull my hand along one its feels like oily leaves on a short vine." (1993, p.65)

What is important about this, for my purposes, is that Michael fails to visually experience the object he discerns via his sense of touch, and hence there is some discord between his visual and tactile modalities. Michael's sensory systems effectively generate a contradiction every time he tastes something. Thus, synesthetes exhibit a certain amount of inter-modal incoherence – a breakdown in the normal pattern of connections between the parts of experience.

On the basis of this sort of evidence I think it would be reasonable for single-track theorists to concede the inter-modal case. However, there is still the matter of intra-modal

---

[70] This is a possible motivation for the single-track theories of Churchland and Baars. In particular, an executive system of the kind Baars envisages is well suited to ensuring the coherence of consciousness, because the process whereby contents gain access to the global workspace, before being broadcast, is a competitive one. This competition guarantees that only "compatible" or "consistent" contents can enter consciousness (see above). So, for Baars, the process of being brought to consciousness, and the process of being made part of a coherent experience, are one and the same.

[71] Of course we know, roughly speaking, how a choir manages to maintain musical coherence: each singer integrates the sound she hears with an internal musical plan of some kind, and produces an appropriate sequence of notes in response. But the point here is that *each and every singer* is capable of disrupting the choral harmony – there is no final arbiter on matters musical, beyond the agreement that can be secured between the singers.

coherence (e.g., the spatial binding of visual properties, such as colour and texture, within a single phenomenal object). Single-track theorists might justifiably resist the view that intra-modal coherence is a merely contingent feature of consciousness. They might argue that the adaptive value of coherence *within* modalities is so great, surely evolutionary pressures will have conspired to guarantee it. Dennett (1991b) may be right to reject the idea of a single, central theatre where everything comes together, but perhaps there are a number of neighbourhood theatres, one for each distinct mode of experience, which ensure intra-modal coherence.

This line of argument is not completely implausible, but in the end it too falls foul of the empirical evidence. Consider, for example, the peculiar visual phenomenology experienced by subjects recovering from damage to striate (visual) cortex. The various elements that normally co-exist in fully integrated visual experience (i.e., colour, shape, depth, and motion), individually "reappear" according to a fixed sequence:

> At first the patient will see pure motion (usually rotary) without any form or colour. Then brightness perception returns as a pure Ganzfeld – a uniform brightness covering the whole visual field. When colours develop they do so in the form of 'space' or 'film' colours not attached to objects. The latter develop as fragments which join together and eventually the colours enter their objects to complete the construction of the phenomenal object. (Smythies 1994, p.313)

The disharmonies that occur during this recovery sequence are striking. Colour and form are not initially bound together in coherent objects: objects appear fragmented; colours are free-floating. The existence of these kinds of experiences clearly militates against an approach that seeks to make intra-modal coherence a necessary feature of consciousness.

One concern about these cases I've cited is that they are both pathological, in some sense. This is clearly true in the case of those recovering from damage to striate cortex. Synesthesia, on the other hand, is experienced (to some degree) by about one in a hundred thousand people, and is not the result of damage (it is an inherited condition). Such numbers suggest we'd better not treat phenomena like synesthesia as peripheral when we frame our theories of consciousness. At any rate, pathological or not, both these cases undermine the view that a mechanism which brings contents to consciousness must at the same time guarantee their coherence. Damage to such a mechanism, if it were to destroy coherence, ought to prevent conscious experience from arising at all. Cases of incoherent experience thus create a problem for the single-track model of consciousness, but are just what one would expect to find on the multi-track model. This is not to deny that consciousness typically *is* coherent; it is, rather, a pervasive but contingent feature of experience, and so must *not* be necessitated by the mechanisms (whatever they are) that bring contents to consciousness. Insofar as a single-track model is posited in order to guarantee coherence, it is gratuitous.

Round one goes to the multi-track theorists. But there is still the matter of *explaining how* the coherence of consciousness arises when it does occur. And it is clear that phenomenal experience typically exhibits a good deal of coherence, both inter-modal and intra-modal. In our daily experience we sometimes only have one source of information regarding external objects: we hear the bird, but we can't see it; we see the ball (on the roof), but we can't feel it. In these cases we don't expect our various modes of experience to be in complete accord; their objects, being distinct, have no obligation to be in temporal or spatial register. However, very often we have access to information regarding a single object via two or more senses. When it comes to our own bodies, in particular, we are information rich. Thus, as I type these words the sound of my fingers striking the keys is in synchrony with both my visual and tactile experiences of the key-strikes; the location of these same key-strikes, as

revealed in my visual experience, is compatible with their apparent auditory location; and my proprioceptive and visual experiences of hand position are consonant. Inter-modal coherence is pervasive when our senses report on common events or objects. Within modalities we also discover a great deal of harmony among the distinct elements of experience. Vision, for example, provides us with information about colour, shape, depth, and motion. But this information is not free-floating, it comes bound together in coherent phenomenal objects whose visual properties co-vary in a consistent fashion.

Notice that there are two aspects to coherence: 1) *temporal coherence*, as exemplified, for example, in the coincidence of visual, auditory and tactile experiences of a key-strike; and 2) *spatial coherence*, which manifests itself in numerous ways, e.g., we see our bodily parts in positions we feel them, we hear sounds emanating from objects in the direction we see them, we experience colours as confined to the boundaries of their objects, and so on. The first of these presents no special difficulty for a multi-track theory of consciousness, as far as I can see, for it is not implausible to suppose that when phenomenal properties coincide temporally, either within modalities or across modalities, this is entirely due to the simultaneity of their vehicles. Thus when a felt key-strike is temporally aligned with its seen counter-part in experience, we should expect to find distinct, but simultaneous content-bearers corresponding to these two phenomenal objects. Of course, this is pure speculation, since we currently know so little about the timing of information-bearing events in the brain. Nevertheless, it's reasonable to imagine 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. In order that we be able to respond appropriately to rapidly changing local conditions, the various determinants of a behavioural response (visual, tactile, proprioceptive, and so forth) must be brought to bear roughly synchronously, so as not to interfere with each other. Thus, the vehicles of these various kinds of information are likely to *be* synchronous.[72] We shouldn't expect such a "trick with wires" to succeed under extreme stimulus conditions, or in relation to distant events (and indeed it doesn't – think of thunder and lightning), but it would be no surprise to discover that our brains are configured in such a way that local events ordinarily generate perceptual synchronies, both across and within modalities.[73]

Spatial coherence, on the other hand, does create a prima facie difficulty for any multi-track polyphonic theory, because such a theory entrenches what Akins calls the Spatial Binding Problem. With regard to visual experience this can be expressed as follows: "given that the visual system processes different properties of the stimulus at spatially distinct sites, how is it possible that we perceive the world in the spatially coherent manner that we do?" (1996, p.30). A related puzzle arises concerning trans-modal spatial harmony: if there is no single place where the many spatially and temporally distributed neural representations of sensory information are brought together, how is it that we see our bodily parts in positions we feel them, hear sounds emanating from objects in the direction we see them, and so on? This puzzle is simply the (intra-modal) binding problem writ large.

Clearly, simultaneity of vehicles is not going to have much bearing on spatial binding.[74] Single-track theorists take this feature of consciousness in their stride by refusing to identify

---

[72] See Churchland & Sejnowski 1992, p.51 for further discussion of this issue.

[73] This move defuses the so called *Temporal Binding Problem* (see Akins 1996 for discussion).

[74] Although Metzinger (1995), for one, has proposed that synchronisation across neuronal assemblies is the mechanism whereby all forms of representational and phenomenal "wholeness" are generated. It is not obvious how synchronisation can account for spatial coherence, given that the various vehicles which contribute to the representation of a particular object, even if synchronous, are likely to be spatially distinct.

visual experience with the machinations of the visual system; a visual content becomes part of phenomenal experience only when it passes through the executive system to which all conscious information is subject. In other words, single-track theorists adopt the spatial analogue of the explanation I proposed for temporal binding: phenomenal properties are perceived as spatially unified by virtue of being encoded in a single, spatially localised system. A multi-track theorist can't adopt this approach to spatial binding. However, it's *not* clear that this makes the binding problem intractable from the perspective of a multi-track account of consciousness. Indeed, it may be no more than a pseudo-problem generated by adopting what Akins calls the *Naive Theory of Perception*: "the thesis that properties of the world must be represented by 'like' properties in the brain, and that these representations, in turn, give rise to phenomenological experiences with similar characteristics" (1996, p.14). As regards spatial properties, this theory requires that the spatial coherence of visual information "must be mimicked by the spatial unity of the representational vehicles themselves" (p.31). And this surely is a *naive* theory. We don't expect the green of grass to be represented by green-colored neural vehicles. Why, therefore, should we expect spatial properties of the world to be represented by corresponding spatial properties of the brain? So long as the contributing sensory systems represent their common object *as* located in the one place (using whatever proprietary means for the representation of space they have at their disposal), then the experience of object location ought to be inter-modally coherent. And the only intra-modal "binding" we can reasonably expect is a binding at the level of contents. In order for the various phenomenal 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. In Akins' terms they must be content-unified, not vehicle-unified.[75]

This analysis of the spatial binding problem suggests that, yet again, the multi-track polyphonic model of consciousness is superior to its single-track rival. If the single-track model is being proposed in order to address the various binding problems, it's plausibility is thereby diminished, for these problems only arise if we adopt a naive and discredited attitude towards mental representation.

Subject Unity

Before concluding I must consider a further sense in which consciousness appears as a unity. Your experiences do not just occur, they occur *to you*; the multifarious perceptual and understanding experiences that come into being as you read these words are somehow stamped with your insignia – they are yours and no-one else's. It is perhaps this salient dimension that Churchland is really alluding to when he talks in terms of consciousness harbouring "the contents of the several basic sensory modalities within a *single unified experience*" (1995, p.214). I will call this form of unity *subject unity*, although it will emerge that there are several different kinds or aspects of unity collected under this head.[76] The question I want to address here is whether subject unity presents a special explanatory problem for multi-track polyphonic accounts of consciousness.

There are two broadly different ways to approach subject unity: (1) treat it as a distinct phenomenal element (or collection of elements) – a "sense of self" that is typically found among the contents of experience; or (2) treat it as a consequence of there being some "here-ness" or "I-ness" that is *built into* each and every element of experience.

---

[75] This implies, of course, that each element of visual experience, in addition to its non-spatial content, also incorporates spatial information. That is, the basic elements of vision are *color-x-at-location-y*, and so on.

[76] Be aware that what concerns us here is the unity of consciousness at each moment, not the unity of consciousness across time.

There are a couple of obvious candidates for a "sense of self". First, there is the distinction we experience between our bodies, and every other object of our acquaintance. While most objects, including our bodies, are discerned through the external senses, only our bodies are also known to us from the inside; we have a body image – a proprioceptive map – which is neither visual, nor auditory, nor tactile. This singular mode of experience, and the more abstract awareness of a boundary between *self* and *other* that it engenders, constitutes a kind of minimal subject unity. Second, there is the very abstract experience of our ongoing personal narrative, the story we tell about ourselves, and to ourselves, practically every waking moment. This narrative, a product of those centres responsible for natural language comprehension and production, comprises a serial stream of self-directed thought (one that non-language using animals presumably lack).

Turning to the second slant on subject unity: it is not easy to make sense of the idea that every element of experience has an "I" built into it. Part of the motivation for this approach is, I take it, to satisfy the intuition that subjectivity is of the essence so far as consciousness is concerned; that there is no sense to be made of phenomenal experiences which lack a subject. While there is something to be said for this idea, it must be admitted that our subjectivity can at times be pretty minimal (think of your experiences just prior to sleep, or on first waking). It is certainly not the case that every element or mode of experience is constantly accompanied by the more abstract kinds of "self" that I described above. But every mode of experience does seem to create a point-of-view, a privileged locus with respect to which each content is "projected". Perhaps we can understand subject unity in terms of the *confluence* of the points of view generated by the various modalities. So long as they 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. Notice that this reduces subject unity to a special kind of coherence – a coherence among the projective loci of the several modes of experience.

These two approaches to subject unity are not mutually exclusive, because we can distinguish between the mere confluence of points of view, and the understanding experience associated with *recognising* this confluence. Such recognition perhaps constitutes a "sense of self" further to those generated by proprioception and the personal narrative.[77] At any rate, I don't really need to choose between these different ways of understanding subject unity, because none of them creates any particular difficulty for a multi-track account of consciousness, as far as I can see. The various senses of self that I've canvassed are but elements of a complex phenomenal whole. They contribute to total consciousness, but are distinguishable from, and independent of much else within experience (as evidenced by the existence of pathologies affecting, respectively, proprioception, the language faculties, and higher thought generally, while leaving the other faculties intact). If anything this analysis of subject unity tends to support a multi-track account. And as for the confluence of points of view: 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 requires a single-consciousness making mechanism (see above).

In the end, however, one may feel that there is something unsatisfactory about all of this; that it has failed to address the real problem of subject unity, what we might call the "boundedness" of consciousness. Hurley expresses the problem admirably when she writes:

---

[77] It is extremely difficult to cleanly separate these different notions of subject unity, and I make no pretensions to have carved the phenomenal field at its joints.

> It seems necessary to distinguish in general between mental states that are *together* within one consciousness and mental states that are not thus together but are in *separate* consciousnesses. How should we understand this difference – the difference between the *togetherness* or *unity* of some mental states occupying a given stretch of time (such as my seeing your face and my hearing my own voice now while I am talking to you) and the *separateness* of other mental states occupying that same stretch of time (such as my seeing your face and your hearing my voice)? (1993, p.50)

Hurley's formulation of the problem implies that mere simultaneity cannot draw boundaries around sets of experiences, and so cannot account for the difference between the "togetherness" of some experiences and the "separateness" of others. Glover makes the same point when he asks us to consider:

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

Simultaneous experiences that belong together, in this sense, are often said to be *co-conscious* (Parfit 1984). Single-track models of consciousness appear to have it all over multi-track accounts when it comes to explaining co-consciousness. If, for example, we assume that in order to enter consciousness a content must bear the appropriate relationship to some kind of central information exchange (a global workspace, say), then it is natural to suppose that two contents will be co-conscious when they enter that relationship simultaneously. It is not simultaneity *per se* that is doing the work here, but the fact that consciousness-making relies on a *single* composer and broadcaster of information; a kind of hub around which the various elements of consciousness revolve. Such an account recommends itself because the required communicative relations presumably only obtain inside the one head, and thus it is clear why experiences instantiated in different heads can't be co-conscious (even if simultaneous with one another). We end up with an account of co-consciousness that satisfies our basic intuitions regarding the "togetherness" and "separateness" of experiences.

The difficulty for a multi-track account, in this regard, is that there is no single consciousness-making mechanism to act as a common point of reference for every conscious content. With multiple sites of consciousness-making one might be right to wonder why we don't all have "either a kaleidoscopic and jumbled 'stream of consciousness' or...a case of 'multiple selves' " (Dennett and Kinsbourne 1992, p.234). A couple of responses suggest themselves. Adopting first a deflationary tack, a multi-track theorist might argue that there is actually no requirement to explain co-consciousness. Taking on that burden assumes there is some matter of fact over and above the judgements we make concerning the unity of our experience. There is no doubt that we judge ourselves to be synchronically unified, but presumably I only judge those phenomenal elements to be *mine* which are appropriately connected with *my* judgement faculties. And *your* experiences are not so connected.

A stronger response is to accept that the mechanisms which explain consciousness will not be the same as those which explain the unity of consciousness. Recall that a single-track account can handle co-consciousness (and its converse) so deftly because a *single* mechanism is responsible for bringing contents to consciousness, which mechanism is therefore also capable of drawing appropriate boundaries around conscious experiences. Multi-track theorists clearly can't link consciousness-making to unity in this way. However, that is not to say that informational relations can't have a role in explaining co-consciousness, on a multi-track account. One can only speculate as to nature of the latter. Perhaps the massive communication links between neural networks provide the conditions for co-consciousness

(such that commissurotomy really does produce two cognitive subjects housed in the one brain). On this story, conscious contents will be co-conscious if their (simultaneous) vehicles are linked by the appropriate neural connections, and not otherwise. Numerous theorists have made this suggestion (see, e.g., Lockwood 1989, Ch.6). Alternatively, perhaps the networks responsible for natural language comprehension and production provide a communicative hub for the other modalities. On this story, it is the special role of the language centers in the mind's causal economy that binds simultaneous experiences together. But be clear: this move doesn't make linguistic capacities a necessary condition for consciousness, rather, they become the glue that unites *pre-existing* states of consciousness (which have their origins in sites distributed throughout the brain).

4.3 Whither Unity and Seriality?

Where does all this leave us regarding the unity and seriality of consciousness? I have the rejected the monophonic model, which attempts to ensure the unity of consciousness by treating experience as a single-file stream of contents. This conception of consciousness can only plausibly be maintained if one is prepared to ride roughshod over the distinctions between consciousness, higher-thought and attention. Polyphonic models of consciousness are thus to be preferred to their monophonic counterparts. However, treating consciousness as polyphonic does not exempt one from addressing the unity of consciousness. Indeed, accepting that phenomenal experience has many simultaneous parts renders the problem of unity all the more acute – how can a many be a one? An obvious solution is to trade phenomenal oneness for physical oneness, to treat the many contents of instantaneous consciousness as the burden of a single information bearing vehicle or a single consciousness-making system. But there are difficulties with this solution. Apart from the fact that a single-track model of consciousness is hard to explain from an evolutionary perspective, it is not recommended by the neuroscientific evidence.

I have canvassed another option: a multi-track model of consciousness, according to which each identifiable element of experience has independent origins in the brain. This model becomes a plausible theoretical option when we recognise that the unity of consciousness does not have to be conceived literally in terms of oneness. Phenomenal experience is unified in the sense that it is typically coherent (both intra-modally and inter-modally), and incorporates a sense of self, or a point of view. These angles on unity, far from requiring a single executive system, are compatible with the hypothesis that the brain incorporates a multiplicity of consciousness-making sites, a hypothesis which makes far more sense of the available data, both phenomenological and neurological.

We arrive at the following: consciousness is not one thing; it is a bunch of co-occurrent, albeit tightly co-ordinated things. But if consciousness is not a unity, what becomes of the seriality of consciousness? At the level of individual modes it is clear that there is some kind of serial progression from conscious state to conscious state, as the perceptual processes generate first one content, and then another, in response to changing patterns of input. On the other hand, the neuroscientific and phenomenological evidence make it equally clear that, considered at the multi-modal level, consciousness is a complex amalgam of many semi-independent streams. This parallelism accords nicely with a multi-track model of consciousness. To assert the global parallelism of consciousness is not to deny that focal attention, or high-level thought have a serial aspect. But there is more to consciousness than the contents of high-level thought. What I deny is that the total phenomenal field is dynamically serial.

5

The Dissociation Thesis

In Chapter 3 I proposed that instantaneous consciousness is a complex aggregate state, composed of many independent phenomenal elements, and thus both disunified (in the sense of not being a single, seamless thing) and highly parallel. In order to defend these claims I went on, in Chapter 4, to critique some existing approaches to consciousness. The view that phenomenal experience is monophonic (and hence serial) proved to be unworkable. Among the polyphonic alternatives there is a tendency to suppose that consciousness must be single-track, i.e., that it must depend on a single consciousness-making mechanism or system in the brain. This view appears to be at least partly motivated by a desire to explain the so called "limited capacity" of consciousness: the (putative) fact that we are only conscious of a tiny fraction of the information being generated by our brains at any given instant. Multi-track theorists may well be able to discover some principled way of delivering this feature of consciousness (albeit perhaps not as elegantly or simply as a single-track theorist), but that is not my task here. Rather, I aim to get to the bottom of "limited capacity". I'll show that what lies at the heart of this notion is the *dissociation thesis*: the claim that information can be explicitly represented in the brain without entering consciousness. While, *prima facie*, the empirical support for the dissociation thesis is good (which partly explains its almost universal endorsement by cognitive scientists), it will emerge that most of the available evidence suffers from severe methodological difficulties.

5.1 Limited Capacity and the Dissociation Thesis

One of the attractions of an executive or single-track model of consciousness is that it can deliver a purported feature of consciousness, its *limited capacity*, quite cheaply. In order to see this it will first be necessary to examine some of the discussion surrounding this notion in the psychological literature, since it is subject to some ambiguities and possible confusions.

The idea that there is a limited capacity mechanism associated with consciousness arises in at least three different contexts, as Baars (1988, pp.33-9) points out:

1) In response to selective-attention experiments, in which subjects who are asked to monitor a demanding stream of information seem to be largely unconscious of alternative, simultaneous streams of information.

2) In dual-task paradigms, in which subjects are asked to do two things at once, such as reacting to a short, unannounced auditory tone, while deciding if two sequentially presented written letters are identical; performance on each task tends to degrade due to competition.

3) In studies of working memory, where it has been found that our short-term storage capacity for novel, unrelated verbal items (such as labels for objects and numbers) is quite limited.

Selective attention is best exemplified in the *dichotic listening* studies pioneered by Cherry (1953) and Broadbent (1958). In the most extreme condition subjects receive distinct speech signals in each ear, via headphones. This allows one message to be transmitted exclusively to

the left ear, and a different message to be simultaneously conveyed to the right ear only. Subjects are asked to monitor one or both channels, or to "shadow" one of the streams of speech (i.e., immediately repeat back what they hear). Cherry found that subjects, when shadowing, were only aware of physical aspects of the unattended message. According to their post-experiment reports subjects did not notice when, for example, reversed speech or language changes occurred. Moray and O'Brien (1967) had subjects monitor dichotic lists of digits to detect occasional presentations of letters. Subjects were sometimes asked to attend to one channel only (the relevant message), ignoring the other (the irrelevant message), and sometimes asked to divide their attention between both channels. They found that "performance with the relevant message was relatively better and performance with the irrelevant message was relatively worse in the focussed-attention condition than performance with either message in the divided-attention condition" (reported in Holender 1986, p.5). These results suggest that attention is a scarce resource, which can be voluntarily traded off between channels, but which has a finite, and quite limited capacity.

Posner has conducted extensive studies using the dual-task paradigm (see his 1978 for discussion). In one experiment subjects receive a visual warning signal, followed by a single visual letter. A second letter that may or may not match the first is presented one second later. On some trials an auditory stimulus (a low-intensity, short-duration, white-noise burst) is presented at various points in this sequence. The subjects' tasks are to: (1) report whether the two letters are identical or not, and (2) respond to the auditory stimulus by tapping a key. Posner and Boies (1971), using this procedure, found that reaction times to the auditory stimulus show a dramatic increase during the interval between the presentation of the first and second letters, suggesting that the primary task interferes with the secondary task (reported in Posner 1978, pp.155-6). This result is striking for two reasons: first, because the increase in reaction time occurs before the second letter is presented, and so cannot be accounted for in terms of motor system conflicts; and second, because the observed interference involves distinct modalities. A natural explanation of this effect is that as attentional resources are allocated to the visual recognition task (following presentation of the first letter), the resources available for the auditory task are correspondingly diminished. That is, such results (and there have been numerous studies like this) suggest that there is a single, resource-limited mechanism associated with attention; a mechanism that is sensitive to input from all modalities.

Finally, studies examining immediate recall of unrelated, novel items, such as words, objects, numbers or rating categories, reveal that we are capable of storing as few as three or four items without rehearsal, and only marginally more with rehearsal (Miller 1956, Newell & Simon 1972). Miller, for example, claims that when it comes to making absolute judgements along a single dimension (be it saline concentration, loudness, pitch or whatever) our capacity for transmitting error-free information is limited to around 2 to 3 bits (i.e., 4-8 rating categories) (1956, pp.83-6).[78] And it is a commonplace that working memory is quickly overwhelmed with intermediate results in the performance of mental arithmetic (e.g., multiplying two or three digit numbers), or in a novice game of chess. It thus appears that working memory (or short-term memory) is quite limited; again there is evidence of a limited capacity system associated with the current focus of attention.

It is now fairly standard to draw a direct connection between these limited capacity phenomena and the nature of consciousness. For example, Stillings et al. claim that "the

[78] This limitation is usually treated as a working memory limitation, because the rating categories are learned during the judgement trials. Far more information than this can be transmitted (over the period of a judgement trial) when the rating categories have been committed to long-term memory. A person with absolute pitch, for example, can identify around 50 to 60 different pitches (5 to 6 bits) without error.

contents of working memory and the focus of attention roughly coincide with the contents of consciousness" (1995, p.42); Baars tells us that "conscious capacity does appear to be quite limited, as shown both by the selective attention experiments and by the limitations of short-term memory" (1988, p.85); Posner and Klein believe that consciousness is related intimately to "the operation of a limited capacity mechanism sensitive to input from all modalities" (1973, p.21); and Mandler wants "to restrict the concept of consciousness to events and operation within a limited capacity system, with the limitation referring to the number of functional units or chunks that can be kept in consciousness at any one point in time", and takes this system to be "modality independent" (1975, p.237).

There are two principal claims here:

1) consciousness lines up with attention and working memory, so evidence for limitations in the latter is evidence for corresponding limitations in the former, and;

2) this evidence suggests that consciousness itself is a modality independent, limited capacity mechanism of some kind.

Despite the almost universal acceptance of these claims among cognitive psychologists, I think we should be wary of too quickly acquiescing in the prevailing climate of opinion. In what follows I'll argue that the case for the limited capacity of consciousness is, at the very least, overstated, and then attempt to expose the more fundamental source of this doctrine.

To begin with, there is clearly some connection between consciousness and working memory: the contents of the latter (such as the digits of a novel phone number) must all pass through consciousness, as it were, both during initial storage, and in the subsequent control of action (such as dialling). But it is important to realise that the contents of immediate experience and working memory do not coincide. Mandler reminds us that: "[consciousness] is not a memory system – it does not involve any retrieval. What is in the momentary field of consciousness is not remembered, it is psychologically present" (1975, p.237). This seems right; rehearsing a telephone number (at least when using the verbal/auditory modality) involves recalling the digits in strict sequence – there is no moment when the whole number is present to consciousness. It is working memory, not instantaneous consciousness, that contains the sequence in its entirety. Of course, what this tends to suggest is that instantaneous consciousness is even more limited than working memory (thus the temptation to "roughly" align consciousness with working memory). But I believe it would be a mistake to conclude this. By and large, studies of working memory have examined the constraints on verbal memory. And while moment by moment verbal consciousness may contain only a single chunk (or at best a few chunks) of data, it is reasonable to suppose (as I argued in Chapter 3) that consciousness as a whole is much richer than this. Instantaneous visual consciousness, in particular, contains far more data than our powers of description permit us to convey.[79] Similar comments apply to the other

[79] Baars rejects this claim, on the grounds that: (1) when presented with an arbitrary number of small unrelated visual objects, subjects have difficulty estimating their number in a single glance (Mandler & Shebo 1982); and (2) we scan even a coherent scene with rapid saccades, and so "the complex scene is not necessarily in perceptual consciousness at any one time: we accumulate it over many serial fixations" (1988, p.85). With regard to the first of these, I suggest that the limitation here resides not so much in visual experience itself, but in our capacity to extract reportable data from experience (I'll have more to say about this issue below). And with regard to the second: it is reasonable to suppose that the purpose of saccadic eye-movement is to allow the brain to *build up* a complex visual scene by numerous fixations of the eye's highly sensitive foveal region. Moreover, while it is legitimate to distinguish between the focussed centre of visual experience, and the relatively unfocussed periphery, it is not legitimate to treat peripheral vision as completely uninformative. In fact, peripheral vision enlarges the compass of visual experience to such an extent, that in a single fixation we can get quite a robust impression of, say, a crowd, or some other extended object.

sensory modalities. What's more, these various modalities, in addition to the current contents of understanding experience, are all simultaneously present in consciousness. It is a gross misrepresentation of the nature of consciousness to reduce it to the current contents of working memory, or to the contents of verbal experience.

It is similarly mistaken to identify consciousness with focal attention. The selective attention experiments suggest that within modalities there is some kind of limitation associated with the more abstract levels of processing. One can be aware of two or more simultaneous conversations, but it is very difficult to render more than one stream of speech intelligible. If one attempts to divide one's attention between two conversations then grasp of each is degraded relative to an exclusive focus on either one of them, just as Moray and O'Brien discovered with respect to their letter detection task. Similar remarks can be made about modalities other than the verbal/auditory (which has been the primary target of selective attention studies to date). Thus, within any given modality, there are demonstrable limits to the capacity of focal attention, limits which we can now quantify, to some extent. Be that as it may, it is arguable whether this observation tells us much about the bounds of consciousness. As Neisser (see Chapter 4), and more recently Mangan (1993), have pointed out, besides the phenomenology of focal consciousness, there is a discernible phenomenology of the periphery or "fringe". The totality of experience within any given modality includes both focal and peripheral elements, and thus clearly exceeds the scope of focal attention. Moreover, it is not unreasonable to suppose that the limits on focal attention (within modalities) are, properly speaking, limits on the capacity to extract more abstract information from that which is already in experience. For example, entering a room full of people engaged in conversation, one is at first aware only of the totality of vocal sounds in the room. But when a particular conversation comes into focus, the unattended talk is not entirely masked; it may even tend to dominate one's experience (particularly in a noisy environment). It is as though special-purpose processing resources come on line, adding detail, but not obscuring the existing phenomenology. This suggests that the resource limitations applicable to attention don't apply to consciousness as a whole, because the contents of focal attention amount only to the more abstract parts of a richer total experience.

What of the claim that consciousness depends on a limited capacity mechanism that is not specific to any sensory modality? The dual task results have clearly pushed theorists in this direction, since attentional interference effects appear to cross inter-modal boundaries. It may be that there is an alternative explanation of the reaction time increases (such as motor pathway inhibition due to pre-emptive competition for control of the motor systems), but for the sake of argument I'll accept the standard interpretation of these results, namely: that attention involves some kind of common, limited capacity resource, a resource which can be accessed via any modality. On the assumption that the contents of consciousness coincide with the contents of attention, what we end up with is an image of consciousness as some kind of informational bottleneck, or central information exchange. I have already discussed some reasons to be wary of identifying experience with the contents of focal attention *within* any given modality. What is being suggested here restricts consciousness even further, because it identifies phenomenal experience with a *single* limited capacity resource for which the various modalities must compete. However, remember that the experimental studies which give rise to these claims – the dual task studies, in particular – depend on verbal instructions, and hence bring high-level linguistic/symbolic capacities on line. Such capacities are likely to be engaged not only while initial training takes place, but *also during task execution* (in the form of instruction rehearsal, and other high-level focussing mechanisms). No doubt there is something to the claim that linguistically mediated thought is limited by its serial nature, and good reason to regard language as a kind of amodal centre for the coordination and control of disparate sources of information. So, by definition, tasks that call on our linguistic capacities depend on a limited capacity system. But as I have

repeatedly stressed, there is no good reason to align consciousness with the activities of such a system. To do so is to restrict consciousness to the most abstract parts of attention (the verbal/symbolic focus), or, putting this another way, it is to regard phenomenal experience as co-extensive with the attentional component of higher-thought.[80] The conflations of consciousness with attention and higher-thought come together here.

Minds as Communication Channels

To bring these criticisms into clearer focus it may help to examine more closely the origins of the view that human information processing capacity is subject to quite severe limitations. We need look no further than Miller's seminal paper "The magical number seven…" (1956). Miller argues that, for the purposes of certain kinds of studies, we can regard human beings merely as communication channels. An important property of any communication system is the correlation between input and output, because this determines how much useful information can be transmitted by the system:

> If you will now imagine a communication system, you will realize that there is a great deal of variability about what goes into the system and also a great deal of variability about what comes out. The input and the output can therefore be described in terms of their variance (or their information). If it is a good communication system, however, there must be some systematic relation between what goes in and what comes out. That is to say, the output will depend upon the input, or will be correlated with the input. If we measure this correlation, then we can say how much of the output variance is attributable to the input and how much is due to random fluctuations or "noise" introduced by the system during transmission. So we see that the measure of transmitted information is simply a measure of the input-output correlation. (1956, p.82)

Humans receive inputs in the form of various kinds of stimuli, and generate outputs in the form of behavioural responses (both verbal and non-verbal). It is possible to regard the correlation between the variance in the stimuli and the variance in the responses as a measure of the information transmitted by the human system. This is typically how absolute judgment experiments are interpreted. Such studies measure the capacity of subjects to track uni-dimensional stimuli. For example, Pollack (1952) assigned numerals to a collection of tones which varied only with respect to frequency over the range 100 to 8000 Hz. He exposed subjects to sets of tones (from 2 to 14 in number) drawn from this collection. Each set was presented in random order (presumably after a training period) and subjects were asked to identify the tones by number. If the sets were limited to two or three tones then subjects were able to remember the identification scheme, and never made mistakes. As the number of tones increased, mistakes became more frequent, and at 14 tones mistakes were very frequent. Careful analysis of the results shows that as the amount of information in the presentation sets increased (from 1 to 3.8 bits) the correlation between the variation in the stimuli and the variation in the responses approached a maximum (an asymptote) of about 2.5 bits. Miller interprets this study as providing a measure of the *channel capacity* of the observer: "it represents the greatest amount of information that he can give us about the stimulus on the basis of an absolute judgment" (ibid.). That is, absolute judgment experiments tell us something about the transmission characteristics of humans considered as communication systems.

Experiments like this are striking because they provide such a concrete handle on human communicative capacities. But notice that they are only able to do so because of the

---

[80] Understanding experience itself has a focus and a periphery: the *substantive* topic, and the many *transitive* feelings of relation which attend it, as William James might put it (see Chapter 3).

extremely impoverished stimuli to which subjects are exposed. Employing restricted stimuli makes it possible to track the correlations between input conditions and the limited kinds of response variance that short-term memory and human language allow. When the stimulus environment is richer it becomes extremely difficult to determine precisely which features of the input are contributing to the behaviour of the system, i.e., it becomes difficult to keep track of the relationship between input and output. One might think that by severely limiting input we give a false impression of the human information processing system. By creating an artificially restricted environment one risks losing sight of two things: (1) the undoubted fact that the brain *adds* information in the course of processing input, it is not a straight-through system[81]; and (2) the enormous richness and complexity of the data our brains represent at each instant. What I'm suggesting is that the analogy with communication systems is not very apt, because a communication channel, unlike the human brain, does not construct representations of the world. Its job is to faithfully transduce signals from one medium (say, soundwaves) to another (say, electromagnetic radiation), with a minimum of distortion or information loss.[82] The human system not only adds information (in the form of tacit assumptions about the perceptual environment) in order to generate its representations, but it has multiple sources of input (it is multi-modal), most of which themselves appear to be multi-channel (the representation of colour, for example – even if we ignore its spatial aspect – involves at least three independent dimensions: hue, saturation and luminance).[83] The crucial disanalogy with communication systems concerns the fact that, when it comes to generating output, these many channels of data come into existence more or less simultaneously, but only the information which is most salient in the light of the system's current needs and interests will actually have behavioural effects. There is a very wide base of internally represented information from which attentional mechanisms determine what will be subject to further processing, and what will be acted upon. It is thus not appropriate to infer from output limitations (including short-term memory limitations) to the existence of corresponding limits in the representational resources of the brain or the contents of consciousness.

Rather than view consciousness as a limited capacity system associated with working memory and attention, I think we should view it as a broad based representational resource, upon which the mechanisms that generate the very abstract representations (and phenomenology) associated with higher-thought can draw. In the ascent from environmental stimuli to the abstract, amodal representations that are responsible for our most complex behavioural responses (including, especially, our verbal ones) there is undoubtedly some kind of processing bottleneck. This bottleneck arises because of limitations in our capacity to extract (and store) abstract, readily reportable information (such as information about numerosity, absolute value on some uni-dimensional scale, etc.) from the rich ground of concrete experience. The motor end of the system acts as a second bottleneck, since the nature of our bodies severely restricts the kinds and amount of information we can transmit. But to associate either of these bottlenecks with total consciousness is a mistake. The vast majority of conscious representations are rapid,

---

[81] This is the message of poverty of the stimulus arguments. See, for example, Chomsky 1959 and 1980.

[82] This second medium is usually one that is capable of transmitting a signal over long distances, and is chosen so as to maximise the throughput of information.

[83] Of course the difficulty is to get a handle on all this richness experimentally, because historically the only means of assessing instantaneous experience has been via the observation of external behaviour. Thus, the tendency to adhere to the time-honoured and commendable scientific practice of holding fixed all but a limited number of variables. I don't really have any beef with this methodology, except insofar as it is assumed to have some significance concerning the extent of the representational resources, and in particular the conscious representations, that the mind is capable of bringing to bear from moment to moment.

mandatory responses to environmental stimuli, which are generated *prior to* the mechanisms of higher-thought. So far as most experience is concerned, the limited capacity bottleneck revealed in absolute judgement and dual task studies is a *post* consciousness phenomenon, and tells us little about the extent of the information, conscious or otherwise, that the brain encodes at each instant. And the limitations in our capacity to *transmit* information, via speech or other forms of action, has even less to teach us about the representational resources of instantaneous consciousness.

The Dissociation Thesis

Having said all this, there's an obvious sense in which we can agree that the contents of instantaneous consciousness are limited: they are *finite* – no information processing system has an unbounded capacity. But, given that it's so obvious, this is presumably *not* what cognitive psychologists are getting at when they refer to the "limited" capacity of consciousness. Talk of a limitation is presumably intended to mark a useful contrast of some kind. The question, then, is: limited by comparison with what?

Miller and Buckhout claim that the primary function of consciousness is *selective*: "[we] are constantly swimming through oceans of information, far more than we could ever notice and understand; without some effective way to select what is important, we would surely drown"(1973, p.55). They go on to provide some examples of what they have in mind: as one views a regular grid of dots, different groupings will present themselves from moment to moment; similarly, on any particular occasion of viewing a Rorschach ink-blot test, some potential interpretations will stand out, others will not; the steady train of clicks produced by a metronome is usually heard as a rhythmic pattern, but for some this pattern will be in common time (4/4), for others it will be in triple time, and so on. In other words, "all perceptions are…selected from an assortment of alternative perceptions we might have had instead" (ibid., p.57). This "selective" approach of consciousness provides some kind of handle on the capacity of consciousness, since it's predicated on the view that there's an enormously large pool of information – "far more than we could ever notice and understand" – from which the moment by moment contents of consciousness are drawn. By contrast with this pool the capacity of consciousness looks decidedly limited. However, notice that the use of "information" here is ambiguous between: (1) the *signals* in the environment to which we are capable of responding (the information *out there*); and (2) the contents of the representations that our brains actually generate (the information *in here*). Read the first way, the comparison between the contents of instantaneous consciousness and the "oceans of information" through which we swim, strikes me as being utterly banal. *Of course* we only consciously represent a vanishingly tiny proportion of all the things we might experience at any given moment. Not only are the vast majority of signals in our immediate environment either too low energy or too poorly directed to be transduced at an appropriate sensory surface, but, of those that achieve transduction, a great many are subject to rapid interference, and don't lead to content fixations of any kind.

Of far greater theoretical significance is the contrast between the totality of information that our brains encode at each instant, and the proportion of this information that actually enters consciousness. The widely accepted view is that the brain represents *far more* information than ever achieves consciousness. And it is this I think most theorists are really getting at, either explicitly or implicitly, when they refer to the "limited capacity" of consciousness. It is by comparison with the vastness of the *unconscious* – the information stored in long-term memory, the information-processing basis of automatic skills, and so on – that the contents of consciousness appear limited. If consciousness has a selective function, then it is presumably to select what is most salient from this enormous quantity of unconscious information.

It goes without saying that in order to adopt this line one must accept the existence of the *cognitive unconscious*, that is, one must accept that information can be represented unconsciously in the brain. However, we need to tread carefully here, for the notion of mental representation is subject to a good deal of ambiguity – there are at least three distinct styles of representation that we can reasonably ascribe to the brain (see Section 2.2). Much of the information encoded in the brain is inexplicitly represented (i.e., it is tacit or potentially explicit). Such information is certainly unconscious, yet I don't think it is this information that theorists generally have in mind when they want to contrast the limitations of consciousness with the much larger capacity of the unconscious. It is the respective *processing* capacities of the conscious and the unconscious that are generally the focus of interest (see Baars 1994). And while it's clear that explicitly encoded information has direct causal relevance to mental processes, it's not clear that the same can be said for information that is merely potentially explicit. Thus, a contrast between conscious and unconscious capacities properly revolves around the information that is *explicitly* encoded in the brain.

A robust limited capacity claim begins with the view that the vehicles of explicit representation in the brain divide into two classes: those whose contents are conscious and those whose contents are unconscious, and goes on to assert that the former are vastly outnumbered by the latter. This way of putting things makes it clear that there are really two theses here, which I'll glorify with names:

1) *The Dissociation Thesis*. Information can be explicitly represented in the brain without entering consciousness – explicit representation and phenomenal experience are dissociable.

2) *The Limited Capacity Thesis*. Only a small fraction of the informational contents explicitly tokened in the brain at each instant is conscious.

The dissociation thesis is so named because it asserts that conscious experience and (explicit) mental representation can come apart or be *dissociated* – not every explicit content is conscious. The limited capacity thesis goes beyond the dissociation thesis, because it makes a claim about the *relative proportions* of conscious and unconscious information explicitly tokened in the brain, specifically: that the vast majority of the information explicitly tokened in the brain at each instant is unconscious. Even so, the limited capacity thesis depends on the dissociation thesis in two important respects. First, as a matter of logic, one can't assert the former without implying the latter. To claim that only a small fraction of the contents explicitly tokened in the brain is conscious, one must clearly accept the *existence* of unconscious, explicit mental representations. Second, and more significantly, it is only by virtue of being expressed in these terms (in terms of explicit representation) that a limited capacity claim has any teeth, because it thereby acquires genuine implications for the causal basis of cognition (see my comments above). Thus, it is the dissociation thesis that is foundational here. And it is this foundational thesis that I want to carefully consider in what follows.

At the beginning of this section I remarked that many theorists are attracted to single-track models of consciousness because they can provide a very straightforward explanation of the limited capacity thesis. Baars, for example, considers it important that his model be consonant with this thesis (1988, pp.84-6). And his model clearly delivers limited capacity, since of all the explicit information generated by the brain, only that which gains access to the recurrent system centred on the thalamus actually becomes conscious (see Chapter 4 for discussion). This result generalises to any single-track model which, like Baars', assumes competition among informational contents for access to the executive system, or which regards the carrying capacity of this system as physically limited in some way. What lies at the heart of this explanatory virtue is the manner in which single-track models handle the

dissociation thesis. Such models distinguish between explicit contents that are conscious, and those that are unconscious, in terms of a single consciousness-making mechanism. And such a mechanism will invariably act as a bottleneck through which only a limited amount of information can pass at each instant.

The remainder of this chapter will be devoted to taking a closer look at the dissociation thesis. In Section 5.2 I'll examine the origins of the thesis, and go on to provide a thumbnail sketch of the unconscious, as traditionally conceived. I'll then proceed, in Section 5.3, to raise some methodological concerns about the empirical work that is typically offered as evidence for the thesis.

5.2 Origins of the Dissociation Thesis

There are three distinct kinds of reasons for the almost universal acceptance of the dissociation thesis by cognitive scientists: (i) historical; (ii) empirical; and (iii) computational. The last of these concerns the dominant role of classical computational theory of mind in contemporary cognitive science, and the kinds of computational resources that classicism is able to bring to bear on the explanation of intelligent behaviour. I'll address this issue in the next chapter. For now I turn to the historical and empirical sources of the dissociation thesis.

Reasons Historical

I propose to make only a few brief remarks about the historical origins of the dissociation thesis. Late last century, when psychology was first becoming a scientific discipline, it was widely held that the proper subject matter of psychology is immediate conscious experience. The structuralists, led by Wundt and his student Titchener, hoped to identify the basic elements or "atoms" of thought (which they took to be sensations) and the laws governing their combination and interaction. They believed that thinking must always involve imagery of some kind, and saw no reason to consider unconscious processes. Their method was experimental introspection, a highly disciplined procedure in which trained observers reported their experiences under controlled conditions. Using this method the structuralists catalogued on the order of 40,000 elementary sensations. However, the synthetic part of their program, aimed at discovering the laws according to which these basic elements are combined in complex forms of thought, never really got off the ground. (Dellarosa 1985)

Structuralism as a research strategy, and introspection as a method, were eventually abandoned due to factors both internal and external to the program. Internal tensions arose when a group of German psychologists based in the city of Wurzburg flaunted the analytic part of the program, and began to study conscious thought processes directly – Wundt had decreed that only simple mental phenomena, such as reflexes and sensations, could be studied in the scientific laboratory. Again the method of choice was controlled introspection. Subjects were asked to free-associate on chosen words, or answer various kinds of questions, and provide detailed reports of their thought processes (see, e.g., Kulpe 1964). It was a major upset to the structuralists, and the cause of some controversy, when a consistent pattern of negative results began to emerge: many subjects reported that they were not aware of any images accompanying their associations. Such "imageless thought" suggested that a psychology confined to studying the contents of immediate experience would, in the end, be unable to provide a satisfactory account of thought. (Mayer 1983, pp.11-3)

External pressures came to bear with the work of Freud and Von Helmholtz, who both challenged the view that the mind is transparent to introspection. Freud divided the mind into three systems: the *Unconscious*, the *Preconscious* and the *Conscious*. The Conscious contains just what we experience at a given moment, and is assumed to require no further characterisation. The Preconscious and the Unconscious, on the other hand, consist of

"subterranean" beliefs and desires –intentional states lacking any phenomenal properties. The two systems differ in this respect: elements of the Preconscious are capable of becoming objects of consciousness (although presently unconscious), while elements of the Unconscious are blocked from consciousness. Crucially, Freud held that both sorts of states can affect our (conscious) thoughts and behaviour, while not generating any phenomenology.[84] Von Helmholtz, a physicist and physiologist who was involved in early work on neural transmission rates, argued that vision depends on unconscious inferences. Such inferences, because they give rise to basic perceptual experiences themselves, are presumably not accessible to introspection. (Johnson-Laird 1993, p.15)

We thus find in Freud and Von Helmholtz the idea of states in which intentional and phenomenal properties come apart, and an important explanatory role for the unconscious. Since Freud conceived an unconscious populated with propositional attitudes (beliefs, wishes etc.), and von Helmholtz spoke in terms of unconscious "inferences", it is pretty clear that both theorists were in effect advocating the dissociation thesis – they were modelling unconscious representations on those explicit mental states of which we are conscious. Their work was a direct challenge to the explanatory adequacy of structuralism – since it fails to make any reference to the unconscious – and an implicit criticism of the method of introspection.

The structuralists faced pressure of rather a different kind with the rise of methodological behaviourism. Behaviourists were also critical of introspection as a research method, but not because it failed to reveal the true sources of thought. Rather, they held that any reference to unobservable mental phenomena, conscious or otherwise, is a breach of good scientific practice, and that mentalistic idioms ought to be completely eliminated from psychological theorising. In particular, behaviourists regarded conscious experience as neither a fit study in its own right, nor an appropriate guide for theory development. They sought to transform psychology into a science of behaviour; a science with the sole aim of discovering laws that govern the relationship between environmental stimuli and the behavioural responses they evoke.

There is no need to here rehearse the fate of this research program, nor the subsequent rise of cognitive psychology.[85] What is important, for my purposes, is the way in which mentalism emerged from the forty year hiatus imposed by behaviourism. Behaviourist strictures about consciousness dovetailed neatly with earlier qualms concerning the introspective method, to produce an emasculated form of mentalism – a mentalism strangely uncomfortable with consciousness, and with introspective reports (except insofar as these serve as a guide to the unconscious basis of cognition). It is thus no surprise to find Fodor asserting that "Practically all psychologically interesting cognitive states are unconscious..." (1983, p.86). This is a neat theoretical inversion: contrary to the nineteenth century tradition, in which conscious experience was taken to be the subject matter of psychology, it is *un*conscious states and processes that have become the focus of attention.

Reasons Empirical

A great many cognitive phenomena are today interpreted in terms of the dissociation thesis. Insofar as they are successful, cognitive explanations couched in its terms provide *indirect* support for the thesis, in the form of a (presumed) inference to the best explanation. In

---

[84] Psychoanalysis focuses on the problem of aberrant behaviours and thoughts, but healthy individuals are no less subject to the effects of unconscious states, according to Freud.

[85] See Dellarosa 1985 and Johnson-Laird 1993 for brief discussions, or Gardner 1985 for a more extended treatment.

addition, a significant number of studies are widely held to provide good *direct* evidence for unconscious representations. I will discuss both these kinds of empirical support, which can be gathered together under four heads: *memory*; *learning*; *perceptual processing*; and *reasoning*.

Memory Discussions of memory are hampered by the fact that this term is applied to an enormous range of situations in which present behaviour exhibits sensitivity to past experience. Memory is said to be involved when we recite a poem or song, when we recall childhood experiences, and when we ride a bike. The last of these is an example of skilled behaviour, and is often taken to involve a distinct memory system that stores *procedural* knowledge, as opposed to the *declarative* knowledge we access when recalling a person or place, a fact or an episode. I'll have more to say about procedural memory shortly. Here I want to focus on declarative memory. Consider the process of recalling a familiar telephone number. The numbers arrive in their proper order, one digit at a time, and usually quite rapidly. At each moment we only have one, or perhaps a few digits before us, which quickly move aside to make room for the next in the sequence. There is a kind of "window" of consciousness, in which the numbers are briefly framed before moving out of sight again. (This effect is more pronounced when one recalls memorised text.) Given this experience of recall, it is quite natural to suppose that a telephone number is stored in long-term memory as some kind of list – an ordered collection of symbols, with a distinct symbol for each digit, and perhaps a label or marker at the beginning of the list. The retrieval process becomes a matter of locating the relevant list, and activating its elements in sequence. Of course, this story presupposes the dissociation thesis, since each consciously apprehended digit is assumed to have a distinct symbolic counterpart residing in the unconscious.

Learning Two kinds of considerations arising from the study of learning have contributed to widespread acceptance of the dissociation thesis. First, there is the phenomenon of automatisation, which occurs during the acquisition of motor or cognitive skills. When we begin learning a difficult task, such as typing, or playing a musical instrument, it is commonly observed that a great deal of high-level attention is required for task execution. With typing, for instance, one must initially attend to the placement of each finger. After skill acquisition, it is possible to respond to whole words or phrases in the text, without laboriously working through the individual letter-to-keystroke correspondences. Moreover, with practice even the typing of long words becomes quite automatic (which is not to say that it is entirely unconscious). We observe here the well known "chunking" of data that occurs when multiple instructions are somehow combined, and appear to the conscious observer as a single entity.

Similar remarks apply to a whole range of skills, including cognitive ones, such as chess. Novice chess players tend to be quite slow (and poor) at making decisions, as they consciously search through the space of possible moves, or employ heuristics that place heavy demands on working memory. While experts also engage in a certain amount of conscious means-ends analysis, they are often able to determine the best response to a given position with only a brief glance at the board. Chess demonstrations in which a single player takes on several opponents simultaneously, spending a few seconds on each move, are a striking illustration of the degree to which decision-making in chess can become automatic. (See Stillings et al. 1995, pp.129-33 for further discussion.) Again the process of attaining expertise seems to implicate data chunking of some kind, with the acquired knowledge being stored in procedural form. Both the typing and chess playing examples suggest that automatisation involves a gradual transfer of the complex, sequential operations of which a novice is conscious, to unconscious processors, which are much more rapid and have a greater information processing capacity. It is natural to assume that very similar symbol-

manipulating processes are going on in novice and expert performance, but in the one case they are slow and conscious, while in the other they are rapid and unconscious.

A second line of evidence, which provides more direct support for the dissociation thesis, concerns the phenomenon of implicit learning. Implicit learning occurs when rules are unconsciously induced from a set of consciously apprehended training stimuli. For example, in the work on artificial grammars first conducted by Reber (1967), subjects are exposed to a set of letter strings generated by a simple grammar. Subsequent performance on a grammaticality test, with novel strings, is often well above chance, suggesting that subjects have learned the rules of the grammar during training. However, since subjects are unable to report the rules of the grammar involved, or give much account of their decision-making, it is reasonable to conclude that both the process and the products of rule induction are unconscious. This is strong *prima facie* evidence for the dissociation thesis, since success on a grammaticality test with novel strings would seem to require explicit representation of the relevant grammar.

Perceptual Processing A great deal of evidence points to the existence of unconscious inferential processes in perception. When a stimulus is degraded, or novel, it is typical for subjects to engage in conscious hypothesis-testing. For example, reading inverted text usually requires careful attention to letter features, and some guesswork based on both semantic and orthographic clues. As Baars points out, this looks like the "conscious analogue of a process that normally takes place quickly, automatically, and unconsciously" (1994, p.6). The processing of ambiguous stimuli also points in this direction. Most words, for example, have multiple meanings, but these usually only enter consciousness one at a time, and in such a way as to cohere with the local semantic context. This suggests that there are unconscious processes of hypothesis-testing at work, which act to resolve linguistic ambiguities prior to sentence comprehension. Visual illusions, such as apparent motion, also motivate the view that unconscious "problem solving" and built-in knowledge of the world play a pivotal role in perception.[86]

Considerations like these are usually interpreted as evidence for the existence of what Stich calls *subdoxastic states*: unconscious, explicit representations "which, though not beliefs, are part of the causal process leading to belief formation" (1978, p.501).[87] A couple of examples will help:

---

[86] For a general discussion of the problem solving approach to perception see Rock 1983. A relevant discussion of apparent motion is Ramachandran & Anstis 1986.

[87] According to Stich, subdoxastic states differ from beliefs in at least two respects. First, they are *deeply* unconscious, i.e., their contents are unconscious *in principle*, whereas it is in the nature of beliefs (*pace* Freud) that they are potentially conscious. Chomsky's grammatical rules are an apt example here: while we all have conscious access to the results of linguistic processing, and can thus readily distinguish grammatical from ungrammatical sentences, there doesn't seem to be any possibility of direct access to the (putative) explicit rule-following involved. Second, subdoxastic states are *inferentially isolated* – the input systems which they inhabit have little or no access to contents tokened elsewhere in the brain. For example, knowledge that one is suffering from a visual illusion seems to be incapable of penetrating the visual system, leading to the celebrated persistence of illusions. Beliefs, on the other hand, are *inferentially promiscuous*; given suitable background beliefs, almost any belief can play a role in the inference to any other. (See Fodor 1983 for similar distinctions.)

1) linguists posit a variety of psychological states to account for our ability to parse and understand sentences – these have the form of representations of phonetic and syntactic structures, and rules for their manipulation, including representations of grammatical information;
2) David Marr's (1982) theory of vision incorporates various "sketches", corresponding to distinct symbolic stages in a processing hierarchy which leads to the production of visual percepts.

Current wisdom has it that subdoxastic states reside in informationally encapsulated input systems.[88] Processing within such systems is thought 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 representation. The culmination of all this unconscious computation is the production of an unencapsulated output: the set of non-inferential beliefs upon which high-level reasoning is based. The contents of sensory consciousness are normally assumed to be associated with some privileged stage in the processing of input. 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. On either story, the vast majority of the representations generated during input processing are taken to be unconscious.

The strong presumption in favour of the dissociation thesis, evident here, is not without experimental support. A number of studies provide quite direct evidence for the presence of unconscious symbols in perceptual processing. I have in mind here work on *dichotic listening*, *visual masking*, and *blindsight*. MacKay (1973), working in the dichotic listening paradigm, asked subjects to shadow (repeat back) the input to one ear, and ignore the input to the other ear. He found that if an ambiguous sentence is presented to the shadowed channel, and a disambiguating word is simultaneously presented in the unattended channel, subjects exhibit a strong bias towards the interpretation suggested by the disambiguating context. Marcel (1983), used a visual masking technique to subliminally expose subjects to written words, and then asked them to decide which of two ensuing words was either semantically or graphically similar to the initial stimulus. Although they could not report having seen the initial stimulus, Marcel found that his subjects were able to perform above chance in these forced choice judgements. The term 'blindsight' was introduced by Weiskrantz to refer to visually guided behaviour that results from stimuli falling within a blind part of the visual field (a scotoma). A number of studies indicate that subjects with ablations to visual cortex can localise flashes of light, or other visual objects, falling within a scotoma (e.g., Perenin & Jeannerod 1975; Weiskrantz 1980). All of these studies, and the many others like them, constitute good *prima facie* evidence for the dissociation thesis. If we don't assume the existence of unconscious, explicit representations in the visual system, for example, it is extremely difficult to reconcile the capacity to make discriminations concerning the visual environment (blindsight), or similarity judgements (visual masking), with the lack of reported phenomenology.

---

[88] Fodor demurs, arguing for the existence of inferentially promiscuous subdoxastic states, such as those that support *modus ponens* when it features in everyday reasoning. He claims that "subdoxastic knowledge of such principles must be accessible to practically all mental processes" (1983, p.85). This is deeply unconscious, yet promiscuous knowledge. If Fodor is right then Stich's two criteria for subdoxasticity (see previous footnote) can come apart.

Reasoning Simple reasoning processes and problem solving also provide some support for the dissociation thesis. The famous Aha! experience – the sudden arrival of the solution to a nagging problem – is a striking demonstration of the significant role played by unconscious activity in our deliberations. It can occur quite spontaneously, with little apparent relationship to immediately preceding thoughts, and in surprising contexts. The mathematician Jacques Hadamard more than once experienced the "immediate appearance of a solution at the very moment of sudden awakening" (1949, p.8). A nice example is related by Poincaré, who had been working on a class of functions later to be called Fuchsian functions:

> Just at this time I left Caen, where I was then living, to go on a geologic excursion under the auspices of the school of mines. The changes of travel made me forget my mathematical work. Having reached Countances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step the idea came to me, without anything in my former thoughts seeming to have paved the way for it, that the transformations I had used to define the Fuchsian functions were identical with those of non-Euclidean geometry. I did not verify the idea; I should not have had time, as, upon taking my seat in the omnibus, I went on with a conversation already commenced, but I felt a perfect certainty. On my return to Caen, for conscience' sake I verified the result at my leisure. (Poincaré, 1982)

Poincaré had not specifically been looking for this identity, but it came to him nevertheless, and at a place and time not of his choosing.

Successful problem-solving appears to consist of three distinct phases: (1) *preparation* – formulating the problem, and making preliminary attempts to solve it; (2) *incubation* – putting the problem aside to work on other projects or sleep; (3) *illumination* – the key insight appears.[89] But, as Baars points out (1994, p.12), this pattern extends beyond creative problem-solving, which is usually conceived as a temporally extended process, to fairly mundane instances of reasoning and thought. Consider being presented with an open sentence, with the task of finding a suitable closure. One is conscious of the incomplete sentence (phase one), and of the needed word, when it arrives (phase three), but the process of retrieval is entirely unconscious (phase two). Likewise, a simple mental search might commence with a verbal thought (such as: *where did I leave those keys?*), followed by the answer (perhaps an image of the coffee table), but what happens in between is utterly opaque. This is what the Wurzburg group discovered when they examined their thought processes. We have conscious access to ends, but no access to means. Miller and Buckhout, following Lashley, remark: "it is the *result* of thinking, not the process of thinking, that appears spontaneously in consciousness" (1973, p.70).[90] What unites simple reasoning with classic cases of problem-solving is the existence of this gap in our experience; a gap which must clearly be bridged by the unconscious. The default assumption is that this bridging implicates many unconscious, explicit representations. Indeed it is hard to see how

---

[89] This is Wallas' analysis of problem-solving (1931). He includes a further phase (*verification*), which is not relevant to my concerns. For further discussion see Mayer 1983, ch.3.

[90] In view of my remarks concerning the nature of higher-thought (see Chapter 3), one might feel that these claims are somewhat incautious. A great many conscious way-stations can occur within a single extended thought process (mental arithmetic is the obvious example). In such cases there is a sense in which we *are* privy to the process of thought, since the steps involved unfold before us in logical sequence. However, even here, we don't have access to the mental activity that takes us from step to step. There are always gaps that must be filled by unconscious processes. Thus, Lashley's claim that "*No activity of mind is ever conscious*" (1956, his italics) is not too far off the mark.

incubation, if it actively contributes to problem-solving, could fail to support the dissociation thesis.

The Unconscious: The Standard Model

A consistent theme emerges from this analysis of the historical and empirical sources of the dissociation thesis. Mainstream theorists treat the unconscious by analogy with what they observe in consciousness: a telephone number stored in memory is conceived as a list of distinct symbols, just like its conscious counterpart; perceptual processing is treated as unconscious hypothesis formation and testing; expert performance is credited to the transfer of consciously executed novice routines to unconscious processors; and unconscious incubation, since it is assumed to be an active phase of thinking and problem-solving, is modelled on conscious formal reasoning. We get an unconscious inhabited by a great many explicit representations, which not only serve as passive vehicles for the storage of retrievable information, but also play the role of causal intermediaries in the unconscious processes that lead from one conscious state to another. On this picture, cognition is subject to a kind of representational and processing homogeneity: explicit representations feature as legitimate vehicles of both conscious and unconscious contents (the representational homogeneity); and unconscious processes are understood by analogy with the causally discrete, structure sensitive operations we observe in much higher-thought (the processing homogeneity).[91]

It is the analogy with conscious thought that drives the dissociation thesis. The dissociation thesis, in its turn, is yoked to the limited capacity thesis, by virtue of the enormous range of complex skills we perform quite automatically, and the enormity of the data-base that supports them. If unconscious activity is like conscious thought, and if unconscious memory is filled "with just those abstract and pictorial symbols that populate our own conscious rememberances" (Dulany 1996, p.181), then it is inevitable that by far the majority of the explicit representations tokened in the brain are unconscious. The homogeneity of cognition reduces consciousness to the tip of an iceberg.

5.3 Doubts About the Dissociation Thesis

The overwhelming weight of current opinion favours a cognitive treatment of the unconscious. On the basis of the evidence I reviewed above it's clear that unconscious processes are ubiquitous in cognition. However, it is one thing to assert that there are a great many unconscious mental processes, and quite another to assert that such processes involve unconscious representations. The robust reality of *conscious* representations is hard to deny. By attending to conscious experience we catch the brain in the very act of representing the world. But it is not out of the question to doubt whether there are any *un*conscious representations, at least of the explicit variety. And indeed, a number of cognitive psychologists have raised such doubts. For example, Dulany (1991, 1996) expressly questions the orthodox commitment to a cognitive unconscious, in which processes analogous to conscious inferencing are assumed to occur. In a critical review of mainstream studies he introduces a very useful distinction between what he calls *contrastive* and *non-contrastive* analyses (1991, 107-11). The former make differential predictions explicitly designed to test for the presence of unconscious explicit representations, whereas the latter simply assume the presence of such states, and proceed to interpret experimental findings in that light. In what follows I will use this distinction in order to classify and criticise the putative evidence for the dissociation thesis.

---

[91] Dulany (1991,1996) and Smolensky (1988) make similar points.

Non-contrastive Analyses

Of the lines of evidence discussed in Section 5.2, a good many turn out to simply assume that which they appear to support. Discussions of automatisation, and problem-solving, in particular, appear to suffer from this defect.

The discussion of unconscious incubation in problem-solving and reasoning has, until recent times, been very much driven by anecdotal evidence deriving from the diaries and memoirs of scientists. This is cause for concern, given the notorious unreliability of human memory, and our tendency to reconstruct the past to suit present needs or theoretical agendas.[92] Notably, Poincaré's famous descriptions of his thought processes were far from disinterested, since he had quite definite ideas about the nature of unconscious incubation. He regarded this phase of problem solving as a period of intense activity, in which unconscious ideas "flash in every direction…like the molecules of gas in the kinetic theory of gases" and whose "mutual impacts may produce new combinations" (Poincaré 1982, p.393). Thus "sudden illumination [is] a manifest sign of long, unconscious prior work" (ibid., p.389). Given this theoretical agenda it is hard to take Poincaré's accounts at face value. Anecdotal evidence has, however, been difficult to replace with firm evidence gleaned from laboratory studies. Researchers have not only in many cases failed to observe incubation effects, but such effects as do appear are generally difficult to replicate.

An exception is the work of Smith and Blankenship (1989, 1991) who have been able to reliably replicate incubation effects. Their studies involve remote associate tests, in which subjects must discover a single word that conjugates with each of three (seemingly unrelated) words to form common words or phrases. For example, the triple (WATER PICK SKATE) constitutes a single test item, for which the solution is ICE. Smith and Blankenship found that they were able to induce incubation effects by fixating problem solving. This involves priming subjects with distracters: information that is likely to lead away from a correct solution. In the example the distracters (bath choose board) were used. Incubation effects were examined by re-presenting unsolved (fixated) test items either immediately after the initial test or after a period of incubation. During the incubation period subjects were given demanding tasks to perform, so as to prevent continued work on unsolved problems. In this context an incubation effect amounts to a "greater improvement in solving initially unsolved problems when retesting occurs after a delay rather than immediately following the initial test" (1991, p.65).

Smith and Blankenship found that all fixation manipulations were effective at reducing initial test scores, and that mean improvements in test scores following fixation were always greater with an incubation period than without. These results are significant because they provide a reliable method for observing and studying incubation effects. However, of greater significance, in my view, is the finding that *no incubation effects were seen when problem solving was not deliberately fixated*. In other words, among subjects who were not exposed to misleading distracters it made no difference to scores on a retest whether an incubation period was involved or not. A very natural explanation of this result, and one that Smith and Blankenship advance, is that the observed incubation effects are the result of a memory retrieval block, which is created by the distracters used to generate fixation. Such a block prevents access to the problem solution, but as time passes its effects gradually wear off. (See Smith 1995a and 1995b for further discussion.) This explanation is supported by the failure to observe incubation effects in the non-fixated condition, and by the further finding (Smith & Blankenship 1989) that memory of misleading information is inversely related to

[92] See Boden 1990, pp.240-4 for a nice discussion of this point.

incubation effects. It is also far more parsimonious than an explanation in terms of unconscious work conducted during the incubation interval.

It must be admitted that these studies do not demonstrate that fixation is either necessary or sufficient to account for the incubation effects encountered in everyday reasoning, as Smith and Blankenship acknowledge. But what this work teaches us is that, when it comes to explaining the efficacy of an incubation phase in problem-solving, alternatives to unconscious reasoning can readily be imagined. Here are some other possibilities:

1) *physical refreshment* – mental fatigue can interfere with problem solving, so a rest may be just what is required to facilitate progress;
2) *serendipity* – as time goes by it is more likely that the problem solver will stumble across the problem solution, or pick up a relevant cue while working on unrelated material;
3) *forgetting details* – sometimes a problem solver is hampered by concentrating on the wrong level of detail, thus a break can improve performance through the loss of information (this is a variation on the fixation theme);
4) *intermittent work* – one may consciously return to the problem from time to time during an incubation period (this is the kind of episode that Poincaré, for example, may well have forgotten, or been disinclined to report, given his theoretical bias).[93]

These explanations range from the blatantly non-cognitive (physical refreshment), through memory effects of one kind or another (serendipity, forgetting details), to the distinctly cognitive (intermittent work), but none of them requires us to suppose that "unconscious thinking" of any kind occurs during problem incubation. Since there is actually nothing in the way of firm evidence to support the unconscious work hypothesis, and a good many quite plausible alternative explanations, I suggest we should treat incubation effects cautiously so far as the dissociation thesis is concerned.

Very similar remarks apply to the standard treatment of skill acquisition. The common view is that expert control of action and judgment involves much the same kinds of algorithms we employ as novices, except that it's fast, effortless and unconscious. Thus, as Dulany puts it: "the very ubiquity of automaticity, which no one denies, is taken as clear evidence of the ubiquity of unconscious cognitive activity" (1991, p.112). Since explicit representations, if they exist anywhere, are clearly involved in the conscious control of action and judgment, the common view of automatisation acts as implicit support for the dissociation thesis. However, a number of prominent theorists now reject the traditional explanation of automatisation, in favour of more parsimonious accounts. Logan (1988) has proposed an instance theory, which treats automatisation as the replacement of algorithm-based processing with single-step retrieval of past solutions or actions from memory. This is very plausible in simple cases (think of learning to add: children first add single-digit numbers by counting, but eventually learn by rote the sums of all pairs of single digits), and Logan has shown that its applicability is quite general. Chase and Simon (1973), in pursuing the nature of expertise in chess, have proposed that expert players rely on an extensive vocabulary of board configurations (of various grains). Detecting a particular configuration automatically evokes possible relevant moves, from among which the player selects. This hypothesis is similar to Logan's instance theory, because it suggests that algorithm-based processing of the kind found among novice chess-players (looking ahead, following explicit

[93] Further possibilities can be found in Perkins 1981, and Boden 1990 (pp.244-5).

rules of thumb, and so on) is largely replaced in experts with single-step recognition and retrieval. Both accounts agree that, during automatisation, the conscious symbol-manipulations characteristic of early performance drop out, rather than transferring down into the unconscious.

The existence of these plausible accounts of automatisation undermines the rather simplistic view that experts get from point A to point B the same way novices do, only faster, and without being conscious of all the intervening steps. Consequently, any support that accrues to the dissociation thesis on a standard interpretation of automatisation shouldn't be given too much weight.

Contrastive Analyses

Non-contrastive analyses are, in effect, a form of inference to the best explanation. Thus, like any abduction, they are capable of being undermined by the existence of a rival theory with comparable simplicity and explanatory scope. Some more direct evidence would put things on a stronger footing. This is the point of the numerous contrastive analyses to be found in the literature, principally in the study of perception and learning. The most influential paradigms are: *dichotic listening* and *visual masking*, which are reputed to provide good evidence for preconscious semantic processing; *implicit learning*, in which unconscious processes appear to generate unconscious rule structures; and studies of *blindsight*. This last, unlike the rest, is conducted with subjects who have damaged brains (specifically, ablations of striate cortex). And the almost unanimous conclusion derived from these studies is that human cognition implicates a great many representations that are both *explicit* and *unconscious*. Below I present a brief survey of this experimental work, with a view to raising some serious doubts about its methodological credentials *vis-a-vis* the dissociation thesis.

Dichotic Listening In dichotic listening tests subjects are simultaneously presented with two channels of auditory input, one per ear, and asked to perform various tasks. Early work within this paradigm was designed to study the nature and limits of attention (Baars 1988, pp.34-5). Cherry (1953) devised a shadowing task, in which subjects are asked to repeat a stream of speech, word for word, while listening to it. Rapid shadowing is very demanding, and if two streams of speech are presented dichotically, one of which must be shadowed, then the other is experienced as no more than a vague vocal quality. Thus, shadowing is a means of inducing focussed attention in dichotic listening. It was soon discovered, however, that information in the unattended channel can have effects on behaviour. For example, Moray (1959) found that the subject's name, when presented in the unattended channel, would divert attention to that channel, suggesting that unattended information is subject to quite a lot of processing, even if it does not enter awareness. 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. I won't consider the latter here, but see Holender (1986) for discussion and critique.

Lackner and Garrett (1972), and MacKay (1973) have done influential work based on the potential for disambiguation of information presented in the primary (attended) channel by information presented in the secondary (unattended) channel. Lackner and Garrett asked their subjects in a dichotic listening test to attend solely to the verbal input in the primary channel and paraphrase the sentences as they were presented. These sentences contained different kinds of ambiguities (i.e., lexical, surface structural and deep structural), and as they were presented a concurrent disambiguating context was presented in the secondary channel. Lackner and Garrett found that: "The bias contexts exerted a strong influence on the interpretation of all ambiguity types" (1972, p.365). So as to rule out the possibility either

that the subjects were aware that the material they were paraphrasing was ambiguous or that they were conscious of the material in the unattended channel, Lackner and Garrett conducted post-experimental investigations of the subjects' experiences:

> At the end of each experimental session subjects were asked whether they had noticed anything unusual about the material they were paraphrasing, and they were requested to describe as much as they could about the material in their unattended ear. None of the subjects had noticed that the material being paraphrased was ambiguous. 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. One or two disambiguating words were presented in the secondary channel simultaneously with the ambiguous portion of the sentences in the primary channel, but apart from this the secondary channel was silent. MacKay also observed a strong bias towards the interpretation suggested by the disambiguating context (reported in Holender 1986).

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

However, not all cognitive psychologists accept the conclusions typically drawn from dichotic listening studies (see, e.g., Holender 1986). Indeed there is reason to believe that the apparent support for the dissociation thesis generated by this research is an artefact of poor methodology. For example, there is the reliance on post-experiment verbal reports as a source of evidence for subjects' states of awareness during the trials. Nelson (1978) has demonstrated that verbal reports do not provide an exhaustive indicator of conscious awareness, because other tests, such as recognition tests, can detect items not revealed in verbal recall tests, while the converse is not true (reported in Shanks & St. John 1994). Even more problematic is the lack of control in relation to the allocation of attention. In the Lackner and Garrett studies there was no measure of subjects' actual deployment of attention, and Holender's analysis of the experimental protocols suggests that attention could not in fact have been fixed on the primary channel (1986, p.7). While MacKay's use of shadowing did provide a better control of the allocation of attention, it is known that attention can be attracted by isolated physical events in the secondary channel. Mowbray (1964) demonstrated that performance while shadowing word sequences in one channel is dramatically decreased by the appearance of single words in the other, indicating that attention has shifted to that channel. This performance decrement occurs *even when the secondary presentations are not recalled* (reported in Holender 1986, p.5). Most strikingly, in experiments designed to replicate the disambiguation effects, but in which attention deployment was better controlled, such effects did not appear (Johnston & Dark 1982; Johnston & Wilson 1980; Newstead & Dennis 1979). Thus, it is reasonable to conclude that the results obtained by Lackner and Garrett, and by MacKay, were entirely due to uncontrolled attention shifts to the secondary channel, shifts that resulted in brief conscious awareness of the disambiguating context, even if this experience couldn't later be recalled.

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

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

Blindsight Studies Among philosophers probably the best known experimental evidence for the dissociation thesis comes from "blindsight" studies. Weiskrantz coined this term to refer to visually guided behaviour that results from stimuli falling within a scotoma (a blind part of the visual field) caused by ablations of striate cortex.[94] Visual capacity in the absence of striate cortex was first detected in (surgically destriated) monkeys in the last century. But it is only quite recently that this same capacity has been firmly established in human subjects (see Weiskrantz 1986, p.16). For example, a number of studies indicate that subjects with striate ablations can localise flashes of light, or other visual objects, falling within a scotoma, which they indicate by pointing or by verbal distance estimate (e.g., Weiskrantz et al. 1974; Perenin & Jeannerod 1975, 1978; Weiskrantz 1980). There is also evidence that such subjects can discriminate patterns of various kinds. A forced-choice technique has been employed, in which subjects are presented with a succession of stimuli of varying orientations or shapes, and they must choose a pattern (from a range of possibilities provided to them) even when they claim not to see the object. Although the results here are quite varied, with many subjects performing only at chance levels, Perenin (1978) found that some subjects could perform above chance, and Weiskrantz et al. (1974), using three pairs of stimuli, found that each of these two-way discriminations could be achieved, provided the stimuli were large, bright and of sufficient duration. (See Campion, Latto & Smith 1983 for a review of this literature.)

The majority of retinal output in the human brain projects to the striate cortex via a pair of subcortical structures called the lateral geniculate nuclei. It thus appears somewhat mysterious that subjects with striate ablations can retain some visual discriminative or orienting capacity. However, there is a second major visual pathway in the human brain, one that proceeds via the superior colliculus, and has an indirect input to prestriate cortex through the pulvinar. A favourite hypothesis to account for the residual visual capacity found in blindsight subjects, despite their loss of striate cortex, is to regard this pathway and associated structures as representing a second visual system, one capable of supporting object localisation and some degree of pattern discrimination. This interpretation is confirmed by the studies with destriate monkeys. (See Campion et al. 1983 for further discussion.) A principal claim of blindsight research is, therefore, that it provides evidence for a subcortical system capable of giving rise to visually guided behaviour. What has generated all the excitement among philosophers, however, is the further contention that such behaviour can occur in the complete absence of visual phenomenology. Blindsight subjects frequently claim that they can't see anything, and that their answers in the forced-choice discrimination tests are merely guesses. It is this lack of visual awareness that presumably led Weiskrantz et al. to coin the term "blindsight". And it is this aspect of blindsight research that provides evidence for the dissociation thesis. For it is reasonable to suppose that visual judgements are mediated by mental representations: in order for *anyone* to make discriminations concerning the visual environment, some sort of representation of that environment must first be generated. On the further assumption that such

---

[94] For a detailed examination of the phenomenon of blindsight, including both the historical background and more recent experimental developments, see Weiskrantz 1986.

representations must be explicit (given that they are occurrent, causally active states), it appears that the phenomenon of blindsight constitutes evidence for the dissociation thesis.

However, one should not be too hasty here; blindsight research is not without controversy. Campion, Latto and Smith (1983) argue that none of the existing blindsight studies provides adequate controls for light scatter. Furthermore, they claim that it's impossible, on purely behavioural 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" (ibid., p.445). In other words, studies to date haven't ruled out the following, more parsimonious hypothesis: that blindsight phenomena are the result of "light scatter into unimpaired parts of the visual field or…residual vision resulting from spared striate cortex" (p.423). Campion et al. support these claims with a number of experimental studies, in which they demonstrate the covariation of localisation, awareness, and degree of light scatter in a hemianopic subject. Together with the methodological concerns raised above, and the failure to observe blindsight in cases of complete cortical blindness (p.445), these results suggest that a reappraisal of the orthodox interpretation of blindsight studies is in order.

There is thus reason to believe that blindsight depends, in one way or another, on processes mediated by striate cortex. Given that such processes normally lead to visual experience, this is somewhat puzzling, since blindsight subjects putatively have no visual experience of the objects they can localise, and/or identify. However, a solution to this puzzle is not hard to find, because it is with regard to this very issue that blindsight research is most seriously flawed. According to Campion et al. "there is wide disagreement about whether the subject is aware of anything at all, what he is aware of, and whether this is relevant to blindsight or not" (1983, p.435). Many authors assert that their subjects were not aware of any stimuli; others report various kinds and degrees of awareness; and some claim that nothing was "seen", but qualify this by conceding that their subjects occasionally do report simple visual sensations (pp.435-6). The disagreement here is probably partly due to equivocation over the use of terms like "aware" and "conscious" (among the researchers), in conjunction with a failure to ask precise enough questions of the experimental subjects. Wilson, in a study of light scatter, found that "if they are simply asked whether they saw the stimulus, [naive observers] may well fail to report the transient overall glare which is produced by stimuli in blind regions of the field" (1968, p.510). Weiskrantz himself acknowledges this difficulty: patient E.Y., when asked to report what he "saw" in the deficient half of his visual field, "was densely blind by this criterion", but "[i]f 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 the dissociation thesis. But in fact the literature contains a great many reports of experiences that co-occur with discriminative episodes. Consider the comments made by D.B., after performing well above chance in a test that involved discriminating between Xs and Os presented in his scotoma. While D.B. maintained that he performed the task merely by guessing, when pushed by Weiskrantz and his colleagues he qualified this:

> 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 talks of "kinds of pulsation" and of "feeling some movement" in response to moving line stimuli (Weiskrantz, 1986, p.67). And as we have seen, he describes a feeling of "jaggedness" and "smoothness" in response to Xs and Os, respectively.

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

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

For my 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. There is a view that learning can occur completely unconsciously (i.e., without any awareness of the stimuli in the training set), based largely on studies of long-term priming in subjects under anaesthesia. However, what positive evidence there is here is offset by an equally significant body of negative evidence, which leads Shanks and St. John to remark that "it would be premature to conclude from the available studies that unconscious learning [with subliminal stimuli] is feasible" (1994, p.371).[95] 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). So it is not typically the *stimuli* that subjects are unaware of in implicit learning situations. It is, rather, the *relationships between the stimuli* that are held to be unconscious (ibid.).

---

[95] I'll return, in the next section, to research involving subliminal stimuli.

For example, Lewicki (1986) (using what Dulany (1996, p.190) calls the "hidden covariation" paradigm), presented subjects with six pictures of women, whose hair was either long or short, and a personality description in each case. Hair length and personality were correlated in this data. The subjects were then presented with a further set of novel pictures, again showing women with long or short hair, and asked to categorise these women as either 'kind' or 'capable'. Subjects answered correctly to a significant degree, but when questioned "not one subject mentioned haircut or anything connected with hair" (1986, p.138, reported in Dulany 1996, p.190). Clearly these subjects were conscious of the stimuli in the training set, but the suggestion is, not only were they unaware of the correlation between hair length and personality during the learning phase, they also had no conscious access to this information subsequently, despite showing evidence of having induced it from the training set.

As another example, consider the work on artificial grammar learning first conducted by Reber (1967). A typical experiment involves exposure to a set of letter strings generated by a regular grammar (or, equivalently, a set of strings accepted by a finite automaton[96]), which subjects are asked to memorise, followed by a further set of novel strings which they must identify as either grammatical or ungrammatical. Subjects are generally able to perform well above chance on the grammaticality task, yet are unable to report the rules of the grammar involved, or indeed give much account of their decision-making. The standard interpretation of this result is that during training subjects unconsciously induce and store a set of rules. These rules are brought to bear in the grammaticality task, but do not enter consciousness (or, at least, are not reportable). As in Lewicki's experiment, there is *prima facie* evidence here that subjects exposed to training stimuli unconsciously acquire explicit knowledge of the relationships among those stimuli, which information guides subsequent decision-making, even though it remains unconscious.

It may be that the standard interpretation is somewhat incautious, however. Shanks and St. John, in their wide-ranging critique, have identified two principal criteria which implicit learning studies must satisfy in order to establish unconscious learning (in the sense specified above). First, tests of awareness must be sensitive to all relevant conscious knowledge (the *sensitivity criterion*); and second, it must be possible to establish that the information the experimenter is seeking in awareness tests is actually the information responsible for changes in the subjects' performance (the *information criterion*). Unless the first criterion is met "the fact that subjects are able to transmit more information in their task performance than in a test of awareness may simply be due to the greater sensitivity of the performance test to whatever conscious information the subject has encoded" (1994, p.373). And unless the second criterion is met it may be that a task in which performance can be improved by learning I (awareness of which the experimenter tests for), can also be improved by learning I* (awareness of which the experimenter fails to test for) and that, while not conscious of I, subjects actually are conscious of I* (ibid.).

I won't consider the sensitivity criterion in detail here, but just note that a great many studies of implicit learning have relied entirely on post-experiment verbal reports (for example, the Lewicki study discussed above[97]), and this method of assessing awareness is known to be less sensitive than, for example, subject protocols generated during training, or recognition tests (see Shanks & St. John 1994, pp.374-5 for discussion). At any rate, it is the

---

[96] See Hopcroft & 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 $A \rightarrow w\ B$ or $A \rightarrow w$, where A and B are variables and w is a (possibly empty) string of symbols.

[97] It is noteworthy that neither Dulany and Poldrack (1991) nor De Houwer et al. (1993) could replicate the learning effects described by Lewicki.

information criterion that appears to have been most deficient among those implicit learning studies supportive of the dissociation thesis. 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" (reported in Dulany 1996, p.193). Moreover, subjects "reported rules in awareness, rules in which a grammatical classification is predicated of features" (ibid.) 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). Vokey and Brooks (1992) also found some evidence that subjects could remember whole strings, and that similarity to remembered whole strings was employed in grammaticality judgements. Thus, a study that looks only for complex rules, or rules based on whole strings, will probably fail to report the kinds of awareness actually relevant to decisions regarding grammaticality; it will fail the information criterion.

Of particular significance is the finding that when reported rules are arrayed on a validity metric (which quantifies the degree to which these rules, if acted on, would yield a correct classification) they predict actual judgements "without significant residual" (Dulany 1996, pp.193-4); 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" (ibid.). 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)

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

Visual Masking Visual masking is one among a number of experimental paradigms employed to investigate *subliminal perception*: perceptual integrations that, due to short stimulus duration, occur below the threshold of consciousness. It involves exposing subjects to a visual stimulus, rapidly followed by a pattern mask, and determining whether or not this exposure has any influence on the subjects' subsequent behaviour. 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.[98] He found that his subjects were able to perform above chance in these forced choice judgements for stimuli between 5 and 10 msec below the supraliminal threshold. In other words, his subjects were correct more often than not in their similarity judgements, despite the fact that the initial stimulus was subliminal.

[98] 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.

Subjects afterwards reported that they sometimes "felt silly" making a judgement about a stimulus they hadn't seen, but had simply chosen the response (in the forced choice situation) that "felt right".

Marcel takes these results to be highly significant and argues that they "cast doubt on the paradigm assumption that representations yielded by perceptual analysis are identical to and directly reflected by phenomenal percepts" (1983, p.197). Indeed, there is *prima facie* evidence here for the dissociation thesis; when a visual stimulus affects similarity judgements it is natural to assume that explicit representations have been generated by the visual system (especially when it comes to explaining successful graphical comparisons), and Marcel's results 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. These reasons fall under two heads, which I will consider separately.

First, there are methodological grounds for caution. Holender 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 argues that this possibility hasn't been properly controlled for in the majority of visual masking studies (pp.21-2). He 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)" (ibid.). 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). I take the concern here, roughly speaking, to be this: a positive result in a visual masking study is a priming effect that occurs when stimulus durations are below the supraliminal threshold; but statistically significant effects only emerge within 5-10 msec of this threshold, so it's quite possible (in this stimulus-energy domain) that fluctuations in the visual system will occasionally generate conscious events; thus, the (small) degree of priming that occurs may well be entirely due to chance conscious events.

Second, even if we ignore these methodological worries, i.e., even if we accept the visual masking results at face value, there are theoretical grounds for doubting the relevance of small priming effects to an assessment of the dissociation thesis. This is because, in order to account for the observed priming effects, it may not be necessary to suppose that any explicit representations are generated by the visual system. Unconscious semantic priming may be entirely due to computational operations that implicate inexplicit information. It is doubtful whether this explanatory option is available to a classicist, for whom a visual effect on a semantic judgement must surely depend on symbolic output from the visual system (see the next section). But connectionists are in a different position, in this regard, given the alternate account of computation available to them. I'll take up this issue in more detail in the next chapter.

Conclusion

It appears that the empirical evidence for the dissociation thesis is not as strong as it is often made out to be. Non-contrastive analyses, which take the dissociation thesis for granted, are generally open to explanations that don't rely on unconscious, explicit representations, and which are independently plausible. Contrastive analyses, which expressly set out to

establish a role for explicit representations in unconscious processing, appear to be methodologically flawed, in one way or another. Attempts to replicate them under more stringent conditions have often seen the relevant effects disappear, or else prove to be the result of simple, unforseen conscious processes.

In the next chapter I will take up the issue of how one might respond to this situation. In particular, I'll recast the discussion in explicitly computational terms, and examine the options that are available to classicists and connectionists, respectively, when it comes to the dissociation thesis.

6

Consciousness and Computation

It is time to start consolidating the plot. In Chapter 2 I examined the two current contenders for a computational theory of mind: classicism and connectionism. The former treats human cognitive processes as species of digital computational processes. The latter instead seeks to understand mental processes in terms of the operation of neurally-realised PDP networks. Subsequent chapters focussed on developing some desiderata for a plausible theory of phenomenal experience, but did so without addressing the nature of mental processes in any detail. The purpose of the present chapter is to bring these two strands of discussion together, by considering how issues of fine-grained computational architecture impact on the explanation of consciousness.

When cognitive scientists apply computational theory to the problem of phenomenal experience they really only have two options. Either consciousness is to be explained in terms of the representational vehicles the brain deploys; or it is to be explained in terms of the computational processes defined over these vehicles. Theories of the first kind (which I'll call *vehicle* theories) are incompatible with the dissociation thesis, while theories of the second kind (which I'll call *process* theories) implicitly assume it to be sound. In what follows I'll demonstrate that there is a close (and principled) link between the dissociation thesis and the classical computational theory of mind. Consequently, classicists have no choice but to adopt a process theory of consciousness – something most theorists at least implicitly recognise. What has not been widely appreciated, however, is that connectionism has the computational resources to dispense with the dissociation thesis, and hence to support a vehicle theory of phenomenal experience. Connectionism thus opens up a much neglected region of the theoretical landscape for serious exploration.

6.1 The Vehicle/Process Distinction

The question to be considered in this section is: How might the resources of cognitive science be exploited to explain the facts of phenomenal consciousness? Given that computation is information processing, and given that information must be *represented* in order to be processed, a *prima facie* attractive suggestion is that phenomenal consciousness is somehow intimately connected with the brain's representation of information. The intuition here is that phenomenal experience typically involves consciousness "of something", and in being conscious of something we are privy to information, either about our bodies or the environment. Thus, perhaps phenomenal experience is the mechanism whereby the brain represents information processed in the course of cognition.

But to *identify* consciousness with mental representation is to assert: 1) that every kind or instance of phenomenal experience is representational; and 2) that all the information encoded in the brain is consciously experienced. And theorists have difficulties with both aspects of this identification. On the one hand, it is commonplace for philosophers to argue that certain kinds of phenomenal experience are not representational (Searle, for example, cites pains and undirected emotional experiences in this context (1983, pp.1-2)); and on the other, it is orthodox in cognitive science to hold that our brains represent far more information than we are conscious of from moment to moment, since our brains store such an enormous amount of data. So sensations, undirected emotions and memories

immediately pose problems for any account that naively 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 instance, there are some philosophers who defend the position that all phenomenal experience is representational (I have in mind here the work of Tye (1992, 1996, forthcoming) and Dretske (1993, 1995)). The general claim is that our phenomenal experience is actually constituted by the properties that our bodies and the world are represented as possessing. In the case of pains and tickles, for example, it is possible to analyse these in terms of the information they carry about occurrences at certain bodily locations.[99] And as for moods and emotions, it is plausible to analyse these as complex states that incorporate a number of more basic representational elements, some of which are cognitive and some of which carry information about the somatic centres where the emotion is "felt" (see, e.g., Charland 1995, Johnson-Laird 1988, pp.372-6, and Schwartz 1990).

Moreover, with regard to the second difficulty, while it is undeniable that our brains unconsciously represent a huge amount of information, there is an obvious modification to the initial suggestion that might sidestep this problem. As we discovered in Chapter 2, it is commonplace for theorists to distinguish between *explicit* and *inexplicit* forms of information coding (see, e.g., Dennett 1982, Pylyshyn 1984, and Cummins 1986). Representation is typically said to be explicit if each distinct item of information in a computational device is encoded by a physically discrete object. Information that is either stored dispositionally or embodied in a device's primitive computational operations, on the other hand, is said to be inexplicitly represented. It is reasonable to conjecture that the brain employs these different styles of representation. Hence an 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*.

I propose to 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 consciousness to emerge from the computational activities in which these representational vehicles engage.[100] Baars' model of consciousness (1988), which I discussed in some detail in Chapter 4, is a representative example. Baars conjectures that consciousness involves competition for control of a global workspace – a kind of central information exchange – among a multitude of unconscious processors. Informational contents become conscious when, as a result of their representational vehicles

---

[99] Tye suggests that "Pains…represent bodily disorders involving tissue damage. Twinges of pain represent mild, brief disturbances; throbbing pains rapidly pulsing disturbances; aches represent disturbances inside the body." (1996, p.297) And, in response to Block's claim that orgasms present a problem for a representational approach to phenomenal consciousness:

> Orgasms are bodily sensations of a certain sort. As such, they can be treated in the same general way as other bodily sensations…In this case, one undergoes intense, throbbing pleasure in the genital region. What one experiences, in part, is that something very pleasing is happening down *there*. One also experiences the pleasingness alternately increasing and diminishing in its intensity. This too is part of the representational content of the experience. One feels *that* the changes in intensity are occurring. There are, of course, a variety of other bodily sensations that are present during orgasm. But they can be treated in the same general way. (1995, pp.268-9)

[100] See, for example, Baars 1988; Churchland 1995; Crick 1984; Dennett 1991; Flanagan 1992; Jackendoff 1987; Johnson-Laird 1988; Kinsbourne 1988, 1995; Mandler 1975, 1985; Newman 1995; Rey 1992; Schacter 1989; Shallice 1988a, 1988b; and Umilta 1988.

gaining access to the global workspace, they are broadcast throughout the brain. The nature of the vehicles here is secondary; what counts, so far as consciousness is concerned, is access to the global workspace. The emphasis is on what representational vehicles *do*, rather than what they *are*. So 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. I shall call any theory that adopts this line a *process* theory of consciousness.

Given that cognitive scientists have available these two quite different strategies for explaining consciousness – one couched in terms of the representational vehicles the brain deploys, the other in terms of the computational processes defined over these vehicles – it is pertinent to ask why so few choose to explore the former path. Why do process theories of consciousness dominate discussion in cognitive science? An answer to this question begins with the recognition that if the explicit representation of information in the brain is to be *identified* with conscious experience, then it's impossible for there to be episodes of unconscious, explicit representation. To assert a vehicle theory is to deny the dissociation thesis.[101] But, as I showed in Chapter 5, there is widespread acceptance of the dissociation thesis today, for reasons both historical and empirical. I particularly have in mind here the experimental work employing such paradigms as dichotic listening, visual masking, and implicit learning, as well as the investigation of neurological disorders such as blindsight. Such "dissociation studies" appear to rule out a vehicle theory.

Another reason for the dominance of process theories of consciousness 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 (see Chapter 2). 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. More importantly, I will argue in the next section, classicism simply doesn't have the computational resources to do without the dissociation thesis. Thus, classicism and the dissociation studies form a perfect alliance. Together they generate 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.

6.2 The Commitments of Classicism

Despite the recent emergence of connectionism, the classical computational theory of mind still shapes much of our thinking about the mind. The dissociation thesis is equally pervasive. Given the crucial role played by computational theory within cognitive science, this situation is unlikely to be a coincidence. What I want to examine in this section is the depth of the bond between classicism and the dissociation thesis. It may be that there are principled grounds for assuming this thesis, given the nature of classical computation. Alternately, there may be a plausible version of classicism that has the wherewithal to dispense with this doctrine. I shall approach the issue via the following question: Is it possible for classicism to regard unconscious representation as all and only inexplicit? The answer, I will argue, is no – classicism is unavoidably committed to the dissociation thesis.

In Chapter 2 I introduced a generic representational framework (due to Dennett) in which to couch the discussion and comparison of computational theories of mind. Of the four styles of representation in Dennett's taxonomy, I argued that only three are germane to

---

[101] Note, however, that denying the dissociation thesis is *not* equivalent to asserting a vehicle theory of consciousness. To reject the dissociation thesis is to assert that every informational content explicitly tokened in the brain gives rise to an element of experience. A vehicle theorist makes the additional conjecture that every element of phenomenal experience corresponds to an explicit representation tokened somewhere in the brain.

explanation in cognitive science, namely: explicit, potentially explicit and tacit representation (implicit representation being a merely logical notion). For classicists to dispense with the dissociation thesis it must be possible for them to explain human intelligence without invoking explicit, unconscious representations. Clearly, this will only succeed if some other style(s) of representation can take up the computational slack, i.e., if the work done by unconscious symbols, in typical classical models of human cognition, can be assigned to operations that implicate some form of inexplicit information. And it will only succeed if there are plausible classical accounts of cognition in which every explicit data structure – every symbol – posited by the theorist lines up with some element of phenomenal experience. The classical theory we are looking for must therefore satisfy two complementary criteria: 1) whenever a symbol in the language of thought is tokened, the content of that representation is phenomenally experienced; and 2) whenever information is causally implicated in cognition, yet not consciously experienced, such information is encoded inexplicitly.

I don't believe a classicist can really contemplate this kind of theory. Difficulties arise in connection with both of its defining criteria. The first – the requirement that every mental symbol contribute to conscious experience – is at odds with standard accounts of memory. A classicist is certainly not committed to the view that our vast long-term information stores are entirely encoded in symbolic (and hence explicit) form. On anybody's story the brain must employ generative mechanisms of some kind, whereby a great deal of what we know is rendered explicit as it is needed. But classicists have characteristically tended to assume that at least some of our stored knowledge (in particular, some of our *declarative* knowledge) is encoded explicitly, in the form of complex symbol structures.[102] This assumption, if adopted by a classical vehicle theorist, leads to the absurd consequence that such knowledge (being explicit) is constantly present to us in consciousness. Accordingly, a classical theorist who seeks to dispense with the dissociation thesis has no option but to suppose that our long-term memories are entirely inexplicit (that is, they are all *encoded* inexplicitly). It seems unlikely that many classicists would be prepared to make this move – it appears somewhat *ad hoc*, from the classical perspective.

The second criterion – that causally active, unconscious information be wholly inexplicit – creates problems of a different kind. 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 the operation of a Turing machine. 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 (see Chapter 2). But, while a Turing machine may have this capacity with regard to a significant body of data, such information can have no influence on the ongoing behaviour of the system until it is actually rendered explicit. Qua potentially explicit, such information is just as causally impotent as the logical entailments of explicit information. In order for information to throw its weight around in a Turing machine it must first be physically embodied as symbols written on the machine's tape; only then can it causally influence the computational activities of the system, once the symbols that encode it come under the gaze of the machine's read/write head.

For my purposes the Turing machine is acting here as an archetype of the physical symbol-processing device. Similar remarks apply to any digital computer, whatever forms of memory and processing it employs. Consequently, when causal potency is at issue (rather

---

[102] Debate (and research) has tended to revolve around the extent and composition of this knowledge base, rather than whether it consists of explicitly or inexplicitly represented data.

than information coding per se) potentially explicit information drops out of the classical picture. Given the version of classicism being considered here, 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, by definition, the data embodied in the primitive operations of such systems. Thus, an unconscious composed exclusively of tacit information would be a causally efficacious unconscious. 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 I'll explain.

Whenever we act in the world, whenever we perform even very simple tasks, it is evident that our actions are guided by a wealth of knowledge concerning the domain in question. This fact about ourselves has been made abundantly clear by research in the field of artificial intelligence, where practitioners have discovered, to their chagrin, that getting computer-driven robots to perform even very simple tasks requires not only an enormous knowledge base (the robots must know a lot about the world) but also a capacity to very rapidly access, update and process that information.[103] So in order to account for problem-solving and reasoning, for example, it is necessary to invoke large bodies of information with a causal role in these processes, but which manifestly do not figure in phenomenal consciousness. To give one simple example, suppose you read in a paper:

> Police have revealed that the victim was stabbed to death in a cinema. Unfortunately, the chief suspect is now known to have been on an express train to Edinburgh at the time of the murder.[104]

On the basis of this information you will probably conclude that the police have the wrong man. However, notice that in reaching this conclusion you have relied on quite a number of beliefs which likely never entered consciousness, such as: a person can't be in two places at once; to stab someone you must normally be in close proximity to your victim; there are no cinemas on express trains to Edinburgh; and so on.

According to the classical vehicle theory under consideration such beliefs *must be tacit*, realised as hard-wired transformations among the various symbol structures implicated in the reasoning process (whose contents are, by assumption, conscious). The difficulty with this suggestion, however, is that, on any half-way plausible account, the unconscious states that mediate the steps of a reasoning process engage in a complex causal economy. While it is possible that there are no explicit rules of inference involved in reasoning, the train of unconscious intermediaries governed by those rules must interact to produce their effects, else we don't have a causal explanation. And the only model of causal interaction available to a classicist involves explicit representations; as Fodor puts it: "No Intentional Causation without Explicit Representation" (1987, p.25). So, either the unconscious includes explicit states, or there are no plausible classical explanations of higher cognition.

On any standard model of perceptual processing the classicist is also inclined to invoke a great many unconscious symbol structures, which act as intermediaries between the immediate products of sensory transduction, and the symbols whose contents constitute our sensory experience (see Section 5.2). Such intermediaries are ruled out by a classical vehicle theory of consciousness. So far as perception is concerned, a vehicle theorist can only grant a causal role to information that is conscious, or that is represented tacitly. This may be treated as an argument against the possibility of a classical vehicle theory of consciousness, but only

---

[103] This becomes particularly acute for AI when it manifests itself as the *frame problem*. See Dennett 1984 for an illuminating discussion.

[104] This example is adapted from Johnson-Laird 1988, pp.219-20.

so long as classicists aren't prepared to adopt non-standard models of perceptual processing. If classicists are willing to accept that there are no interlevels of explicit representation between the transduction of input and the output of sensory systems (in other words, that the processing is "straight-through"), or that, for each of the interlevels of explicit representation that figure in a plausible classical model of perception, there are corresponding elements of sensory experience, then a classical vehicle theory can handle perception. As it happens, Pylyshyn has toyed with the idea that low level vision, linguistic parsing, and lexical access may be explicable merely as unconscious neural processes that "instantiate pieces of functional architecture" (1984, p.215).[105] However, I suspect that most classicists would find both of these options unpalatable.

There is a further problem with this version of classicism: it provides no account whatever of learning. While we can assume that some of our intelligent behaviour comes courtesy of endogenous factors, a large part of our intelligence is undeniably the result of a long period of learning. A classicist typically holds that learning (as opposed to development or maturation) consists in the fixation of beliefs via the generation and confirmation of hypotheses. And this process must be largely unconscious, since our learning typically doesn't involve much conscious hypothesis testing. As in the case of reasoning, this picture of learning requires an interacting system of unconscious representations, and, for a classicist, this means *explicit* representations. If we reject this picture, and suppose the unconscious to be entirely tacit, then there is *no* cognitive explanation of learning, in that learning is always and everywhere merely a process that reconfigures the brain's functional architecture. But any classicist who claims this 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 number of *unconscious symbols*. Consequently, classicists can accept that tacitly represented information has a major causal role in human cognition, and they can accept that much of our acquired knowledge of the world and its workings is *stored* in a potentially explicit fashion. But they cannot accept that the only explicitly represented information in the brain is that which is associated with our phenomenal experience – for every conscious state participating in a mental process classicists must posit a whole bureaucracy of unconscious intermediaries, doing all the real work behind the scenes. In short, classicism doesn't have the kinds of computational resources that would allow it to dispense with the dissociation thesis. Consequently, classicists are precluded from adopting a vehicle theory of phenomenal experience, since such a theory amounts (in part) to a denial of the dissociation thesis. Any classicist who seeks a *computational* theory of consciousness has no choice but to embrace a process theory – a conclusion, I think, that formalises what most classicists have simply taken for granted.

This situation go a long way towards explaining the dominance, within cognitive science, of process theories of consciousness. In conjunction with the apparent evidence for the dissociation of explicit representation and phenomenal experience (see Sect. 5.2) it is little wonder that vehicle theories are virtually absent from contemporary theorising about consciousness. But there are a couple of reasons for thinking that vehicle theories are due for reappraisal. First, a number of theorists have lately raised doubts about the degree to which the dissociation thesis is supported by the available evidence. Experimental analyses which simply take this thesis for granted are generally open to (independently plausible) explanations that don't rely on unconscious, explicit representations. And the various contrastive analyses – experiments expressly designed to establish the validity of the

[105] It is worth noting that Fodor rejects this picture, arguing that such cognitive tasks do implicate processes defined over intermediate explicit representations (see his 1983).

dissociation thesis – appear to suffer from methodological flaws of one kind or another. (See Sect. 5.3.) It is thus no longer obligatory for theorists to assume the dissociation thesis. Second, while the foregoing has established the dissociation thesis as a requirement on classical cognitive science, it has not done the same for cognitive science per se. In the next section I will address this same issue from the connectionist perspective.

6.3 Connectionism and the Dissociation Thesis

If one is intent on adopting a computational vision of the mind, but has reservations about classicism, then there is currently only one place to turn: the connectionist computational theory of mind. In Chapter 2 I developed the contrast between classicism and connectionism, and found that there are significant computational differences between them. Here I aim to show how these differences bear on the explanation of consciousness. In particular, I will establish that connectionism, unlike classicism, *does* have the resources to abandon the dissociation thesis, and thus is not precluded from hazarding a vehicle theory of phenomenal experience.

To abandon the dissociation thesis is to claim that the only information explicitly represented in the brain is conscious information. It is to adopt the hypothesis that every content explicitly tokened in the brain gives rise to an element of conscious experience. As in the classical case, this hypothesis will only succeed if the role of the cognitive unconscious can be entirely taken up by operations defined over inexplicit forms of representation. But is this suggestion any more plausible in the connectionist context, than it is in relation to classicism? I think it is, for while I was able to apply Dennett's taxonomy of representational styles to both classicism and connectionism, these accounts of mind nonetheless exhibit an important representational asymmetry. Whereas potentially explicit information is causally impotent in the classical framework (it must be rendered explicit before it can have any effects), the same is not true of connectionism. This makes all the difference. In particular, whereas classicism, using only its inexplicit representational resources, is unable to meet the demand for a causally efficacious unconscious (and is thus committed to a good deal of unconscious symbol manipulation), connectionism holds out the possibility that it can.

We saw earlier that 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 the particular connection weights and connection pattern of the network. However, recall that this same configuration of weights and connections is also responsible for the manner in which the network responds to input, and hence the manner in which it processes information. This means that the causal substrate driving the computational operations of a PDP network is identical to the supervenience base of the network's potentially explicit information. So there is a strong sense in which it is the potentially explicit information encoded in a network (i.e., the network's "memory") that actually governs its computational operations.

This fact about PDP systems has major consequences for the way connectionists conceptualise cognitive processes. Crucially, information that is merely potentially explicit in PDP networks need not be rendered explicit in order to be causally efficacious. There is a real sense in which *all* the information that is encoded in a network in a potentially explicit fashion is causally active *whenever* that network responds to an input. Thus, a connectionist account of reasoning, for example, need not invoke explicit unconscious intermediaries in order to link one conscious episode with another, in complete contrast with its classical rivals. For example, in the reasoning about the murder suspect described above, various beliefs unconsciously influence the conclusion drawn (a person can't be in two places at once, to stab someone you must normally be in close proximity to your victim, etc.). But a

connectionist need not suppose that such beliefs must be explicit in order to have their effects. A PDP system is capable of exhibiting sensitivity to an enormous amount of data *without* rendering it explicit, by virtue of storing this information (in potentially explicit form) in the very substrate that supports processing (see Section 2.3). For this reason, a set of unconscious beliefs is capable of influencing the course of processing in a PDP system, even while they are all merely potentially explicit.

What is more, learning, on the connectionist story, need not rely on unconscious processes of hypothesis formation and confirmation (at least not insofar as such processes are understood to rely on explicit representations). In a PDP system learning involves the progressive modification of a network's connection weights and pattern of connectivity, in order to encode *further* potentially explicit information. Learning, in other words, is a process which actually reconfigures the potentially-explicit/tacit representational base, and hence adjusts the primitive computational operations of the system. In Pylyshyn's (1984) terms, one might say that learning is achieved in connectionism by modifying a system's functional architecture.

The bottom line in all of this is that the inexplicit representational resources of connectionist models of cognition are vast, at least in comparison with their classical counterparts. In particular, the encoding and, more importantly, the *processing* of acquired information, are the preserve of causal mechanisms that don't implicate explicit information (at least, not until the processing cycle is complete and stable activation is achieved). Consequently, most of the computational work that a classicist must assign to unconscious symbol manipulations, can in connectionism be credited to operations implicating inexplicit representation. Explicit representations feature only as the *products* of unconscious processes. In other words, the inexplicit representational resources of connectionism are rich enough to permit a denial of the dissociation thesis. This is the minimal condition on any vehicle theory of consciousness. Thus, connectionists can feel encouraged in the possibility of identifying phenomenal experience with the explicit representation of information in the brain – unlike classicism, connectionism does appear to have the right computational profile to hazard a vehicle theory of phenomenal experience.

Having established this possibility for connectionism, it is by no means assured that such a theory will prove satisfactory. However, since vehicle theories are all but absent in contemporary cognitive science, it is vital that this much neglected region of the theoretical landscape be opened up for serious exploration. In summary the connectionist account goes something like this. Conscious experiences are stable states in a sea of unconscious causal activity. The latter takes the form of *intra*-network "relaxation" processes that result in stable patterns of activation. Unconscious processes thus *generate* explicit (activation pattern) representations, which the connectionist is free to identify with individual phenomenal experiences, since none is required to account for the unconscious activity itself. The unconscious *process*, entirely mediated by superpositionally encoded data, generates a conscious *product*, in the form of stable patterns of activation in neurally realised PDP networks. What is most distinctive about this account is not the fact that it's a *connectionist* theory of consciousness (a number of theorists have proposed such accounts – see, for example, Phaf & Wolters 1997), but that it's a *vehicle* theory – an approach to consciousness *made possible* by the unique representational and computational features of the PDP framework. In the next chapter I will develop this connectionist vehicle theory in some detail. I'll demonstrate that it's a robust, and defensible alternative to the plethora of process theories in the literature, and one that does justice to the insights about consciousness gleaned over the last several chapters.

7

A Connectionist Theory of Phenomenal Experience

A vehicle theory of consciousness holds that phenomenal experience is to be explained, not in terms of what mental representations *do*, but in terms of what they *are*. Such a theory identifies consciousness with the explicit representation of information in the brain. As I've already indicated there are a couple of reasons theorists have tended to be wary of this hypothesis. First, a vehicle theory of consciousness is incompatible with the dissociation thesis (Section 5.1). But we've seen that the experimental evidence for the dissociation thesis isn't so strong that we may not reasonably entertain its denial (Section 5.3). Second, this approach doesn't appear to fit with the classical computational theory of mind (Section 6.2). Until recently this has put vehicle theories right out of contention, and ensured the proliferation of process theories of consciousness in their stead. Now the situation has changed. Given the potential of connectionist styles of inexplicit representation to account for unconscious thought processes and learning it appears that connectionism has the computational resources to hazard a vehicle theory of consciousness (Section 6.3). Baldly stated the theory is this:

> *phenomenal experience is identical to the brain's explicit representation of information, in the form of stable patterns of activation in neurally realised PDP networks.*

It is the purpose of the present chapter to defend this theory. I begin by examining the origins of the theory, with a view to clarifying and developing its central claim. Then, in Section 7.2, I outline some of the theory's strengths. These include:

1) its ability to do justice to the phenomenological and neurological evidence that support a multi-track approach to consciousness;
2) its explanatory resources with respect to the diversity of sensory experience; and
3) its capacity to account for the various degrees of abstractness exhibited by the elements of experience.

In Section 7.3 I show how connectionism is able take up the explanatory burden carried by unconscious, explicit representations in conventional accounts of cognition, and can thus make good on a denial of the dissociation thesis. In the final section of the chapter I discuss the relationship between phenomenal consciousness and informational access as it emerges in the present account.

7.1 A Multi-track Vehicle Theory of Consciousness

The proposal before us is this: that whenever a stable pattern of activation is generated in our brains, we are consciously aware of a certain item of information; and conversely, that whatever we consciously experience is the result of stable activation being achieved in one or more of the many neurally realised PDP networks in our heads. This connectionist account of consciousness is not completely novel. What sets it apart from other connectionist approaches is: 1) insistence on *stability* as criterial for consciousness; and 2) an explicit focus on the multiplicity of consciousness-making mechanisms implied by a vehicle theory of consciousness.

Theorists involved in laying the foundations of the connectionist approach to cognition recognised a potential role for stable patterns of activation in an account of phenomenal experience. In the very volumes in which connectionism receives its first comprehensive statement (Rumelhart & McClelland 1986; McClelland & Rumelhart 1986), for example, we find the suggestion that:

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

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

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

It is worth pointing out, however, that neither Rumelhart, Smolensky, McClelland and Hinton, nor Smolensky, take the presence of a stable pattern of activation to be both necessary and sufficient for consciousness. Rumelhart et al. don't appear to regard stability as *necessary* for consciousness, for they suppose "that there is a relatively large subset of total units in the system whose states of activity determine the contents of consciousness", and that "the time average of the activities of these units over time periods on the order of a few hundred milliseconds correspond to the contents of consciousness" (1986, p.39). But this implies that "on occasions in which the relaxation process is especially slow, consciousness will be the time average over a dynamically changing set of patterns" (1986, p.39). In other words, stability is not *necessary* for conscious experience, since even a network that has not yet stabilised will, on this account, give rise to some form of consciousness. Smolensky, on the other hand, doesn't regard stable activation to be *sufficient* for consciousness, and says as much (1988, p.13). Consequently, it is not clear that either of these early statements actually seeks to *identify* consciousness with stable activation patterns in neurally realised PDP networks, as I am doing.[106]

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

---

[106] It is important to note that both Smolensky, and Rumelhart et al., develop their proposals in the context of connectionist accounts of thought. In other words, they are interested in *conscious thought*, as opposed to consciousness per se. Smolensky, in particular, focuses on the issue of how conscious rule interpretation might be realised in a PDP system. However, as I emphasised in Chapters 3 and 4, conscious experience encompasses more than the phenomenal concomitants of such (higher) thought processes. It is not clear what these theorists would want to say about simple sensory states of consciousness.

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

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

Lloyd provides a detailed analysis of phenomenal experience, developing the distinctions between sensory and non-sensory, primary and reflective forms of consciousness. He goes on to show how, on the basis of the identity claim above, these various distinctions can be cashed out in connectionist terms (1995, 1996). But, again, Lloyd appears to focus his efforts on activation patterns in general, rather than stable patterns of activity, and so his account in this respect is still at some variance with mine.[107]

Why then have I made *stability* such a central feature of my 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 stabilisation, the activation levels of the constituent processing units of a PDP network are rapidly changing. At this point in the processing cycle, therefore, while there certainly is plenty of activity across the network, there is no determinate *pattern* of activation, and hence no single, physically structured object that can receive a fixed interpretation. A connectionist vehicle theory of consciousness is thus committed to identifying phenomenal experience with *stable* patterns of activation across the brain's neural networks. On this story, a conscious experience occurs whenever the activity across a neural network is such that its constituent neurons are firing simultaneously at a *constant* rate. The physical state realised by this network activity – the complex object consisting of a stable pattern of spiking frequencies in a set of interconnected neurons – that state *is* the phenomenal experience.

One might think that this identification of consciousness with stable patterns of activation in neural networks is inconsistent with the *seamless* nature of our ongoing phenomenal experience. This would be a mistake. Network stabilisations can occur very rapidly; given their chemical dynamics it's possible for real neural networks to generate many stable states per second (Churchland & Sejnowski 1992, Ch.2). Consequently, what from the perspective of an individual neural network is a rapid *sequence* of stable patterns, may be experienced as a more or less *continuous* phenomenal stream. Clearly, assuming my account there must be some limit on the extent to which, say, a moving object in one's visual field will be perceived as tracing out a smooth, continuous path. But in fact the phenomenology of motion perception conforms to this expectation. There is a striking partition of motion related experience that depends on object speed: slow moving objects exhibit what is known as *optimal* (or *beta*) motion – their movement from point to point appears smooth and continuous, and their boundaries remain well defined throughout; rapidly moving objects, on the other hand, generate what is known as *phi* motion – this is figureless motion, which is experienced as a homogeneous, blurred infilling of the traversed space (try moving your pen rapidly back and forth between two points).[108] We might

---

[107] 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 realised PDP networks.

[108] See Kolers 1972, pp.9-10 for these terms, which are normally used to describe kinds of apparent motion. They can, I think, be applied to the perception of motion more generally, with little risk of confusion.

conjecture that phi motion occurs when the visual system reaches the limit of its capacity to rapidly stabilise and restabilise, and so generates a kind of default "moving object" phenomenology in which figural details are neglected.

Another reason for focusing on stable activation patterns is that only stable patterns of activation can facilitate meaningful communication *between* PDP networks, and hence contribute to coherent schemes of action. Such effects are mediated by the flow of activation along connection lines, and its subsequent integration by networks downstream. No network can complete its processing (and thereby generate explicit information) unless its input is sufficiently stable. But stable input is the result of stable output. Thus, one network can contribute to the generation of explicit information in another only if itself in the grip of an explicit token. The message is: stability begets stability. This, I suggest, makes an identification of conscious experience with *stable* patterns of activation particularly appealing, because such patterns have the kind of robust causal role that we intuitively associate with conscious states.

It is important to be aware, however, that in emphasising the information processing relations enjoyed by explicit representations, I am not claiming that these vehicles must have such effects in order for their contents to be phenomenally experienced. This, of course, would amount to a process theory of consciousness. On the vehicle theory I have been developing, phenomenal experience is an intrinsic, physical, *intra-network* property of the brain's neural networks.[109] On this account, therefore, *inter-network information processing relations depend on phenomenal experience*, *not the reverse*. I will have more to say about this issue towards the end of the chapter.

At this point it is timely to consider the second of the features that sets my connectionist account apart from those that precede it. While the emphasis so far has been on the properties of individual networks, it is important to recognise that connectionism is *not* committed to viewing the brain as a single, massive PDP network. That view flies in the face of everything we know about the localisation of function in the brain. Connectionism holds 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. This means that from moment to moment the brain is simultaneously realising a large number of explicit representations. On the vehicle theory of consciousness I am proposing, *each* of these explicit representations – *each* stable activation pattern – contributes to our conscious experience. Consequently, a conscious subject's phenomenal experience at any one moment in time will not generally be the product of a single neural network. It will in fact be a very complex *aggregate* state composed of a large number of distinct phenomenal elements, each of which is generated at a distinct site in the brain.

In Chapter 4 I introduced the distinction between *single-track* and *multi-track* polyphonic models of consciousness, both of which acknowledge that instantaneous consciousness incorporates a multitude of informational contents (primarily because it is polymodal), but which provide different accounts of how this rich experience is realised in the brain. A single-track model of consciousness treats consciousness as the product of a single consciousness-making system or mechanism (see Section 4.2). According to a multi-track model, on the other hand, each of the many discernible elements of instantaneous experience

---

[109] Put this way it is clear that the connectionist vehicle theory of phenomenal experience makes a *type identity* claim, because it identifies conscious mental states with a physically defined class of brain states. Its pedigree therefore extends back to the line initiated by theorists such as Place (1956) and Smart (1959). But this theory puts a new twist on the generic type identity theory, because the high-level physical type it picks out is at one and the same time a *computational* type, specifically, the explicit representation.

arises from a localised, physically distinct structure or mechanism. Hence, phenomenal experience is the product of a *multitude* of consciousness-making mechanisms scattered throughout the brain. In these terms, the connectionist vehicle theory I have mooted is clearly an instance of a multi-track model of consciousness, because it treats phenomenal experience as the sum total of all the many stable activation patterns being generated in the brain at any moment in time.[110] As such, an individual's (global) phenomenal consciousness is simultaneously multi-channelled, multi-modal and massively parallel.

My use of the term 'parallel' here has the potential to create some confusion, given that connectionism makes a quite distinct use of the serial/parallel distinction. The processing in PDP networks is usually said to be "parallel" because it involves the simultaneous activity of a great many interacting units. Multiple local effects determine the outcome of the network relaxation process, and result in a stable state of activation. But there is also a kind of *serial* progression from stable state to stable state, as the network generates first one content, and then another, in response to changing patterns of input. As Rumelhart, Smolensky, McClelland and Hinton put it:

> the "distributed" in "parallel distributed processing" brings a serial component to PDP systems. Since it is patterns of activations over a set of units that are the relevant representational format and since a set of units can only contain one pattern at a time, there is an enforced seriality in what can be represented. (1986, p.38)

This usage is perfectly legitimate, but notice that it is really only an accurate description of what takes place in *single* networks, or small groups of interconnected networks. It is certainly true that large arrays of networks in the brain often stabilise more or less synchronously, but just as often there are independent, asynchronous channels of activity in train. So, in addition to the parallelism of *sub*-network relaxation, the brain also exhibits parallelism at the *super*-network level. This parallelism consists in processes of stabilisation and restabilisation occurring at multiple, independent sites throughout the brain. On a multi-track vehicle theory of consciousness each of these distinct stabilisation processes, if it runs to completion, will result in the creation of a distinct element of experience. The connectionist account of phenomenal experience I propose thus suggests that, at any given moment, consciousness comprises the set of stable activation patterns currently being generated by the brain; and, over a slightly longer time frame, it consists of multiple, parallel *sequences* of stable states.

Before turning, in the next section, to the virtues of the connectionist vehicle theory of phenomenal experience, let me briefly consider a couple of objections to this account. First, there are subjects in deep sleep or coma. One might think that the brains of such subjects, precisely because they are not receiving any fresh input, will have settled into "stable patterns of activation". Yet we don't ordinarily suppose that such subjects are phenomenally conscious. The appropriate response to this worry is that, while the brains of subjects in deep sleep or coma are to a large extent "resting", it is *not* the case that their cortical neural networks have settled into *stable* patterns of activation. A resting neuron tends to fire spontaneously at a random rate (Churchland & Sejnowski 1992, p.53). Consequently, while there is certainly some activity across the neural networks of the brain, even in dreamless sleep, no stable patterns of activation are being generated. According to the theory before us

---

[110] Clearly, none of this really gets off the ground if there is no principled way to individuate networks. One obvious suggestion is to identify networks with groups of units that have the capacity to co-stabilise. An alternative is to look for topological criteria of individuation (e.g., the presence of hidden layers). Such criteria may well be conditioned by the requirement that, in order to count as a network, a group of units must be independently capable of generating motor and/or cognitive responses.

this means that there is no consciousness either, in accordance with the standard intuitions. 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 my account, dreams, just like normal waking experiences, are composed of stable patterns of activity across these networks.

A second objection concerns the brain's control of autonomic processes in the body. Surely there are neural networks associated with these operations which sometimes settle into stable patterns of activity; yet there is no attendant phenomenology. However, while there are certainly neurons associated with autonomic processes, it is not clear whether it is correct to talk of neural *networks* here. To identify phenomenal experience with stable patterns of activation across neurally realised PDP systems, is to claim that the former is a *network* property of the brain. Network properties, patently, are not instantiated by individual neurons, nor do they inhere in groups of neurons lacking an appropriate degree, or form, of connectivity. Autonomic processes are mediated by neurons confined to the spinal cord, brainstem, and neural ganglia. These neurons are generally few in number, lack significant connectivity, and are quite widely scattered through the body. Thus, insofar as the relevant pathways have been mapped out, it is not clear that they ought to *count* as networks. Given my account, it is no surprise that autonomic processes don't generate any phenomenology.

7.2 Some Support for the Theory

The connectionist vehicle theory of consciousness has a great deal to recommend it. Let me now consider some of its most notable features.

(1) *Activation pattern representations have the right temporal characteristics to be identified with conscious states.*

Connection weight representation is suited to the long-term storage of information, since connection weights (which are realised in the brain as variations among the synaptic junctions between neurons) typically have quite long half-lives. By comparison, stable activation patterns are short-lived. They are suited to representing rapidly varying environmental conditions, and short-term goals and targets. Note, however, that by virtue of their stability, activation pattern representations are not *so* short-lived that they fail to constitute a coherent response to the flux of incoming perceptual signals. In this they differ from activation patterns in general, most of which constitute mere way-stations through which some network or other passes as it relaxes into a set of constant firing rates. Given these temporal differentiae it is not unnatural to propose an alignment of conscious states with stable patterns of activation across neurally-realised PDP networks.

(2) *The connectionist vehicle theory of consciousness supports the additive, parallel nature of consciousness.*

In Chapter 3 I argued that instantaneous consciousness is a multi-modal aggregate – a composition of distinct phenomenal elements, and that ongoing consciousness comprises a set of relatively independent parallel streams. I went on to suggest in Chapter 4 that the evidence of phenomenology, in conjunction with what we know about the neural basis of information processing, supports a multi-track model of consciousness. But if this is the case, then the connectionist vehicle theory of consciousness is just the sort of account we need. On this account, phenomenal experience has the complex synchronic structure it does because it consists of a multitude of physically distinct explicit representations generated across the brain from moment to moment. And ongoing consciousness comprises a set of relatively independent parallel streams because it is supported by the activities of groups of more or

less isolated PDP networks.[111] Moreover, on this account, phenomenal experience exhibits patterns of breakdown consistent with a high degree of neural distribution because the very mechanisms that fix explicit contents in the brain are those that generate consciousness. In other words, since it is most natural to interpret the connectionist vehicle theory of phenomenal experience as a multi-track model of consciousness, all the support that accrues to multi-track theories in general (which I canvassed in Chapter 4) also accrues to this theory.

(3) *The connectionist vehicle theory of consciousness has the explanatory resources to account for the great diversity in sensory experience.*

Our conscious experience is extraordinarily diverse, as I hope I was able to convey, to some extent, in Chapter 3. This diversity is most clearly evident in the multi-modal nature of sensory experience, and in the great variety of its parts. One can't really claim to be offering a theory of consciousness until one has, at a minimum, a fairly rich and systematic account of the *similarities* and *differences* among these elements, both within and between modalities.

A striking discovery of experimental psychology is that judgments of qualitative similarity and difference *within* modalities can be codified, in geometrical form, as multi-dimensional "quality" spaces (see, e.g., Land 1977).[112] Consider, for example, our perception of *colour*. Human beings are capable of discriminating at least 10,000 distinct colours. When people are asked to make judgments such as whether one colour is more similar to a second than to a third, whether a colour is between two others, and so forth, over a whole range of colours (judgments which can be scaled using number of just noticeable differences), it turns out that these discriminative capacities map into a three dimensional colour space (Hardin 1988). Each point (or small volume) in this space corresponds to a distinct colour, and distances in the space reflect perceived degrees of similarity between colours. Such spaces can also be discovered for other modalities, and for quite abstract perceptual qualities. A great strength of the connectionist vehicle theory of consciousness is that it has the potential to account for the properties of these *phenomenal* spaces via the *activation* spaces associated with PDP networks in the brain (Churchland 1989, Ch.5, Churchland 1995, Ch.2, Lockwood 1989, Ch.7). An activation space is a vector space associated with a PDP network, each axis of which reflects the activity of a single unit in the network. A point in activation space thus corresponds to a possible pattern of activation over the entire network. If we assume that stable activation patterns in neural networks are identical to phenomenal experiences, then the *phenomenal* relations that obtain between conscious experiences in any one domain can be explained in terms of the *structural* relations between the activation patterns realisable in some network. Colour experiences that are very different, for instance, can be thought to correspond with stable patterns of activation that map onto widely separated points in

---

[111] Clearly, the independence among the parts of experience is a matter of degree. The various sensory modalities exhibit a great deal of independence – a change, or deficit in one seems to have little bearing on the others. Within modalities, however, the story is more complicated. While the various elements of visual experience, for instance, don't normally come apart, we know they are capable of independent existence given the peculiar (and now well-documented) visual phenomenologies of those suffering ablations in the visual cortex. Moreover, even though the more abstract elements of visual experience (such as visual gestalts – see Section 3.1) probably can't exist in the absence of the lower-level elements, they do exhibit a certain "looseness" of fit with these less abstract parts of visual experience. For example, a feeling of familiarity doesn't always attend facial "parsing", and a Necker cube can have more than one interpretation. From the perspective of the connectionist vehicle theory of consciousness these various degrees and kinds of dependence among the parts of experience are amenable to treatment in terms of the brain's inter-network, and inter-system architecture.

[112] This work on *relative* judgments should be carefully distinguished from the work on *absolute* judgments that I discussed in Section 5.1. Relative judgments of the kind described above, unlike those involving absolute judgments, don't require the subject to remember ordered sets of stimuli.

activation space; while points that are near neighbours in activation space correspond with colour experiences that are phenomenally similar.

Let me develop this idea a bit further. According to the connectionist vehicle theory of consciousness each explicit representation generated in the brain is identical to an element of phenomenal experience. To be more precise, each explicit representation is identical to an experience in which the information encoded by that vehicle is "manifested" or "displayed" – that is, the "what-it-is-likeness" of each phenomenal element is *constituted* by the information that some explicit representation encodes. The *differences* in the phenomenology of these elements are explicable in terms of the *differences between the activation patterns* involved. However, because phenomenal differences are both intra-modal and inter-modal, we need to distinguish here between *systems* of patterns and *individual* patterns. Modalities line up with systems of patterns, and a system may be identified with the activation space of the relevant network (or networks). Particular experiences within a modality, on the other hand, are constituted by specific activation patterns – individual members of the system. So *inter*-modal differences (e.g., the difference between an experience of red and the sound of a trumpet) are not to be accounted for in terms of patterns within the one network, but in terms of the structural differences *between* networks. These include both "dimensionality", which is determined by neuron numbers, and "shape", with depends on precisely how these neurons are connected. Both of these features can be brought to bear in accounting for the differences between broad classes of experience. What *unites* colour experiences is that they correspond to patterns of activation in a network with a particular structure (shape and dimensionality). But equally, what *distinguishes* them from experiences of, say, sound, are these same properties; properties which distinguish one neural network from another, and hence, on my account, one *kind* of phenomenology from another. And the same applies, mutatis mutandis, to *intra*-modal differences (e.g., the differences between shape, colour and motion within visual experience). In short, *the pattern is all*.

Some will find this approach to the diversity of sensory experience objectionable. How can mere alterations in stable firing patterns explain the enormous richness and variety of human consciousness? It is certainly a remarkable fact about the brain that this one neural substrate is capable of generating all the different kinds of experience we're capable of entertaining. *Any* physicalist account that attempts to grapple with all this variety tends to have an air of incoherence about it, and the least I can say of my suggestion is that it is no more implausible than any other. In fact, my conception of the relationship between neural substrate and phenomenology is notable for offering the basis of a *principled* story about the manifold nature of sensory experience. Moreover, it finds support in the remarkable experiments involving sensory substitutions conducted by Bach-y-Rita (1972), and reported by Dennett (1991b, pp.338-44). Bach-y-Rita developed prosthetic devices for blind subjects in which very low resolution camera images were communicated to grids of either electrical or mechanically vibrating "tinglers". These grids were then placed on the back or belly of the blind subjects. Even after only a brief training period, as Dennett explains, "their awareness of the tingles on their skin dropped out; the pad of pixels became transparent, one might say, and the subjects' point of view shifted to the point of view of the camera, mounted on the side of their heads" (1991b, p.341). Here, it seems that a pattern of activation generated in a part of the brain normally dedicated to tactile perception is responsible for something that starts to feel (look?) like visual phenomenology.

(4) *The connectionist vehicle theory of consciousness can account for variations in the degree of abstractness exhibited by the elements of experience.*

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, facial familiarity, to highly abstract language-based understanding experiences. Nor should we forget the "fringe" of consciousness, such as the tip-of-the-tongue phenomenon, feelings of familiarity, and the various feelings of relation described by James, which in general fall towards the more abstract end of the scale.[113] It is difficult to render "degree of abstractness" in quantitative terms, but Lloyd (1996) has identified the primary dimensions along which it varies. These are most easily picked out via their end points:

- *sensory experiences* (least abstract) are modality specific, basic (meaning that they are not constituted by or dependent on other elements or experience), and compulsory;
- *non-sensory experiences* (most abstract) are modality independent (because they have no necessary connection with any one modality); non-basic, and voluntary (see Lloyd 1996, pp.65-8).

As I remarked in Chapter 3, the designations *sensory* and *non-sensory* merely mark the extremes of several continua, and most elements of experience lie somewhere between these poles. The experience of melody, for example, is somewhat more abstract than the experience of isolated tones, because it involves perceiving a relationship (or relationships) between a set of notes ordered according to some familiar principles. It thus depends on, and is partly constituted by, those experienced tones. The perception of rhythm is also like this, and is subject to a degree of voluntary control (for example, with a little practice one can learn to impose a variety of rhythmic overlays or interpretations on the ticking of a clock). In general, as one ascends towards the non-sensory (i.e., the most cognitive) elements of experience one finds that these express more and more abstract[114] relationships between or attitudes towards the lower-order elements of experience.

In addition to its capacity to account for the diversity of sensory experience, the connectionist vehicle theory offers a simple approach to this characteristic of consciousness. What we know of neural architecture indicates that the networks of which the brain is composed form a rough hierarchy. Some are very close to the sensory transducers, and receive their principal input from these, others are second-order (i.e., they receive their input from the first layer of networks), and so on. It is natural to suppose, on the connectionist vehicle theory of consciousness, that less abstract elements of experience correspond to stable patterns of activation in lower-order networks, while more abstract elements of experience correspond to activation patterns in higher-order networks. Understanding experiences, in particular, (which incorporate both metacognitive and propositional forms of awareness) presumably correspond to stable patterns of activation in very high-order networks, networks that receive input from many sources, and are thus least modality specific and most subject to voluntary control (possibly via multiple recurrent pathways – see Churchland 1995, p.217). Thus, the variance among the elements of experience with respect to their degree of abstractness is explicable in terms of an underlying physical property of the brain – the hierarchical organization of its constituent networks.

To illustrate this suggestion, consider again the experience of being familiar with a face. Churchland (1995, pp.38-55) reports some fascinating work done on face recognition by Cottrell and colleagues at the University of California, San Diego. This group has modelled

---

[113] See Mangan 1993b for extensive discussion, and Chapter 3 above. Note that Mangan's discussion generally fails to distinguish among some quite diverse kinds of "fringe" experience. Transient (i.e., short-lived) experiences, "fuzzy" experiences (i.e., experiences of ill-defined objects, such as those which inhabit the periphery of vision), and abstract experiences are all important, but inhabit quite distinct subclasses of the fringe.

[114] In the sense of being at once more focussed, and more general – see Section 3.2.

the process of recognising and naming faces using a three-layer, feed-forward PDP network. The input layer of this network consists of a 64x64 grid of units, each of which has a range of 256 different activation levels. This allows the input layer to represent recognisable human faces, in the manner of a grey-scale newspaper image. The hidden layer consists of 80 units, each of which receives input from the entire input array, and sends output to a set of 8 output units. The latter are divided into a group of five, which is used to encode arbitrary numerical labels for specific faces, one binary unit for registering facehood, and two further units which respond maximally to male and female faces, respectively. Cottrell's group trained this net up on 64 photographs of 11 different faces, and 13 photographs of scenes not involving faces. After training the network responded with 100 percent accuracy to the training set, with respect to gender, faceness, and facial identity (by producing the appropriate numerical label), and correctly identified 98 percent of novel photos of people in the training set. More impressively, it was also able to distinguish between face and non-face images with 100 percent accuracy when presented with a set of completely novel faces and scenes, and got the gender of these novel faces correct around 80 percent of the time.

This network acts as a suggestive model for the kinds of neural structures which may be involved in specialised perceptual processes such as face recognition. Although clearly not among the most abstract of mental operations, face recognition is some steps removed from the basic processes of object parsing. Those who suffer with prosopagnosia often remain capable of identifying a face *as* a face (thus, basic form detection mechanisms must still be in place) but have lost the more abstract capacity to *recognise* a face – to see it as *familiar*, with all the wealth of associations that such familiarity implies (see Chapter 3). While the model network devised by Cottrell's group is in many respects neurologically unrealistic[115] it does capture something of the hierarchical nature of visual processing. Moreover, it suggests a place for feelings of familiarity in the scheme of things. If we treat the input layer of this model as corresponding to the lower-order networks responsible for form detection, then it's quite natural to suppose that feelings of familiarity are produced in the brain analogue of the hidden layer. This layer generates reduced dimensionality activation pattern representations of the input data, by virtue of detecting correlations hidden in the input array. Such correlations can be shared by distinct input patterns, and an appropriately trained network will pick these out – hence, the capacity of Cottrell's network to recognise the one individual in a number of different photos (including photos it hasn't been trained on). An analysis of the hidden layer activation patterns involved here reveals that the vectors associated with a familiar face tend to cluster tightly in activation space (Churchland 1995, p.54). It is this clustering that enables layer three to respond so robustly when the network is exposed to someone familiar: the structures responsible for encoding abstract facial features (the stable activation patterns in layer two) are physically very similar to one another. The same is not true of vectors associated with presentations of an unfamiliar face. These are likely to be scattered in activation space, and hence physically rather dissimilar. One might conjecture that it is stable patterns of activation falling within these "familiarity zones" in hidden-unit activation space (the regions where vectors cluster for familiar faces) that correspond to feelings of familiarity.[116]

---

[115] The role of the first layer of Cottrell's network is occupied, in a real brain, by a good part of the visual cortex, thus placing the networks responsible for face recognition at least five synaptic steps downstream of the retina (Churchland 1995, p.51). The third layer is nothing more than a schematic for those systems involved in identifying the gender of a face, putting a name to a face, and so on. In addition, Cottrell's network is exclusively feed-forward. The visual systems are known to incorporate a great many recurrent connections.

[116] What one really wants here is an interactive network model. Familiarity is a form of abstract experience that can be present in some instances, absent in others. If familiarity is generated in a higher-order recurrent network

More generally, the neurological evidence to date is not inconsistent with the view that abstract experiences are the products of higher-order networks. It would not be unreasonable to suppose that it is in part the brain's hierarchical structure that invests consciousness with its complex sensory/non-sensory structure. Again a significant feature of human experience emerges naturally from the connectionist vehicle theory of consciousness.

7.3 The Dissociation Thesis Revisited

So far I have set out the principal features and strengths of the connectionist vehicle theory of consciousness. However, an issue that still looms large is the need to take up the explanatory burden carried, in conventional accounts, by the unconscious. My account, since it *identifies* consciousness with the explicit representation of information in neurally-realised PDP networks, involves a rejection of the dissociation thesis. Consequently, I am committed to the view that human cognition can be explained without recourse to unconscious explicit representations of any kind. In Chapter 5 I made use of the distinction between *contrastive* and *non-contrastive* analyses (Dulany 1991) as a means of systematising the kinds of evidence that bear on the dissociation thesis. Contrastive analyses are specifically designed to provide evidence of explicitly represented unconscious information. Non-contrastive analyses, on the other hand, only support the dissociation thesis by way of an inference to the best explanation. I earlier suggested some alternative explanations for the phenomena that fall under the purview of the non-contrastive studies, and cast doubt on a great many of the contrastive analyses (Section 5.3). Here I reverse this order of exposition, first reconsidering one particularly influential contrastive study (Marcel's work on subliminal perception), and then returning to the non-contrastive phenomena. My aim is to establish that connectionism has the potential to make good on a denial of the dissociation thesis.

Subliminal Perception

Marcel (1983) used a series of visual masking experiments to investigate so-called subliminal perception: the occurrence of short duration, unconscious stimuli that appear to have some effect on behaviour. In these experiments subjects were exposed to a short-duration visual stimulus, rapidly followed by a pattern mask, such that the initial stimulus never registered in consciousness. Marcel found that his subjects were able to perform above chance in forced choice judgements for stimuli between 5 and 10 msec below the supraliminal threshold (see Section 5.3 for further details). Following Holender (1986), I raised some doubts about the methodology of visual masking studies. Suppose, however, that we take Marcel's results at face value. The question I want to raise here is whether such results create a serious problem for the connectionist vehicle theory of consciousness. Marcel reasons that, in order for the initial stimulus to affect ongoing cognitive processing (for it to be "cognitively effective"), a representational vehicle of some sort has to be generated and manipulated in the system. And if so, it is clearly a vehicle that gives rise to no associated phenomenal experience.

But there is some question as to whether the visual masking results, *when interpreted from within the connectionist camp*, unequivocally lead to the conclusion that Marcel draws. In particular, one should always be mindful of the *style* of computation employed by connectionist systems. Once a connectionist network has been exposed to input, it takes a certain amount of time for it to "settle" or "relax" into a stable pattern of activation (though here, of course, we are talking in terms of milliseconds). Initially the network will oscillate rapidly between highly dissimilar firing patterns. However, just prior to stabilisation the

---

responsible for encoding abstract facial qualities of some kind, then failure to stabilise would correspond to the absence of that phenomenology.

network is likely to converge on some small region of activation space, such that moment to moment variations in its activation pattern become negligible. At this point, even though an activation pattern representation hasn't yet been generated (since the network isn't completely stable), there is enough consistency of output to allow some meaningful communication with other networks in the system.

With this in mind I can offer the following explanation of Marcel's experimental findings. Because of the short duration of the initial stimulus, and because of the subsequent pattern mask, there is not enough time for a stable pattern of activation (and thus an explicit representation) to be generated in the relevant visual networks. But there is enough time for some significant effects to fan out from these networks to other parts of the brain. Recall that in Marcel's experimental studies the initial stimulus was cognitively effective only when its duration was marginally below (i.e., no more than 10 msec below) the supraliminal threshold. At this juncture the visual network's activation profile is only likely to vary marginally from moment to moment. Thus, although an explicit representation hasn't been generated, and therefore no conscious experience produced, enough processing has occurred for there to be meaningful effects on those networks responsible for our language capacities. It is this meaningful "pre-conscious" communication that accounts for the accuracy of the subjects' responses in a forced choice situation.

The general point here, then, is that cognition, from the connectionist perspective, is a good deal more holistic than classicism allows. A certain amount of information processing takes place in a connectionist network prior to the production of an explicit mental representation, and it is reasonable to suppose that this processing has some coherent inter-network effects.[117] If this is right, then I can conclude that the visual masking results, interpreted from within the connectionist framework, do not show that phenomenal experience and explicit mental representation are dissociable.

Non-contrastive Analyses

Memory The role of memory in cognition is both ubiquitous and diverse. Researchers regularly distinguish among *explicit*, *implicit*, *procedural*, *declarative*, *long-term* and *working memory* (see, e.g., Anderson 1995), to name a few, although it is highly unlikely that all these forms of memory are subserved by distinct computational or neurological mechanisms. My interest here primarily centres on declarative memory, and in particular on working memory. Long-term memory, be it explicit or implicit, episodic or factual, is almost universally assumed to depend on changes at the synaptic level (see, e.g., Rose 1993). Since it appears that the encoding of data by weight modification is superpositional in nature, information stored long-term creates no particular difficulties for a vehicle theorist. There is little reason to suppose that such information takes the form of explicit unconscious representations. At any rate, long-term memories are certainly not stored as stable patterns of activation in neural networks.

Working memories, on the other hand, are plausible candidates for storage in explicit form. Working memory is a form of short-term memory[118] which appears to contain a

---

[117] Having conceded this, it would not do to exaggerate the significance of these effects. Remember, first, that they only arise (if in fact they do) when the initial stimulus duration is *just marginally below the supraliminal threshold*, and second, that such effects are *small* primings in a forced-choice situation. One might think that the limited nature of these priming effects would discourage researchers from taking the idea of explicit unconscious information too seriously.

[118] Anderson claims that the idea of a distinct short-term memory system – a limited-capacity store through which information must pass in order to be laid down in long-term memory – has now fallen into disrepute (1995, pp.171-4). Working memory is not "short-term" in this sense, but merely decays rapidly. Information doesn't have to spend time in working memory in order to be stored more permanently.

number of distinct components: an auditory loop; a visuospatial loop; a verbal (articulatory) loop; and possibly some other components (Baddeley 1986). Within each of these loops it is possible to maintain items of information for about 1-2 seconds without rehearsal, or much longer if they are continually refreshed. An unfamiliar phone number, for example, is preserved in the articulatory loop from the time you look it up in the phone book until you begin to dial (or until you consciously rehearse the number). Intermediate results generated during mental arithmetic are likewise maintained in the visual or verbal store while other steps of the calculation are performed. It is during the period between conscious rehearsals that an item is stored in working memory.[119] And it is tempting to suppose that such an item takes the form of an explicit unconscious representation. This is a guiding presupposition of both classical and connectionist models of working memory. For example, both Zipser (1991), and Dehaene and Changeux (1989), appear to assume that working memory depends on the maintenance of stable levels of activation among interconnected sets of units.[120]

However, it is far from clear that recourse to unconscious symbols constitutes the only viable explanation of working memory. One possibility that is consistent with the connectionist vehicle theory of consciousness is the idea that working memory is realised by a cellular mechanism (such as a transient synaptic modification, or a change in firing threshold) with the same temporal dynamics as working memory itself (Churchland & Sejnowski 1992, p.304). The role of such a mechanism would be to (briefly) alter the activation profile of the system. A content stored by these means, though not matched by a stable pattern of activation during storage, would have an increased probability of being realised by such a pattern (and hence entering consciousness) while the transient mechanism was in place.

A second possibility concerns the capacity of recurrent networks to maintain a circulating signal for some period of time. In addition to its capacity to converge on a *point attractor* in activation space, which corresponds to a stable activation pattern, such a network is capable of tracing out a closed path in activation space known as a *limit cycle* (see Churchland 1995, pp.97-104 for discussion). When a recurrent network enters a limit cycle the instantaneous firing rates of its constituent neurons vary from instant to instant, so there is no point at which network activity stabilises. Nevertheless, by virtue of tracing out a closed path in activation space such a network is restricted in its range of possible activation states. One can imagine a memory mechanism based on placing a recurrent network in a limit cycle, which would act to keep it in the vicinity of one or more point attractors representing distinct items of information. Input from a gating mechanism of some kind (á la Zipser 1991) or from external networks, could then be used to direct the network towards one of these attractors. At this point the relevant item of information would enter consciousness, even though it was at no stage explicitly represented prior to doing so. So far as I know no-one has proposed this kind of memory mechanism, but it has the advantage that it fits the phenomenology of working memory, i.e., the partially involuntary timing of item recall. And it also coheres with what we know about the activity of cells in prefrontal cortex thought to implicated in working memory. As it happens such cells don't exhibit stable spiking rates, rather they fire at higher than background, but *varying* rates (Goldman-Rakic 1990, 1992).

Whether either of these suggested mechanisms is a satisfactory approach to working memory is anyone's guess at this stage, but they are at least viable options. And for the

---

[119] I'm assuming that it wouldn't be useful to regard information as being in working memory when it is actually conscious, although there is a strong temptation to talk that way on these time scales.

[120] Zipser's model is discussed in Churchland & Sejnowski 1992, pp.301-4.

vehicle theorist they both offer the advantage of not requiring that informational contents be explicitly represented in order to be stored in short-term memory. The jury is still out on the mechanisms responsible for working memory. What I have established here is simply that the available data doesn't force us to the conclusion that short-term active memory involves the explicit representation, via stable patterns of activation, of the data being stored.

Learning The primary phenomenon I must contend with here is automatisation, because this form of learning is so frequently interpreted in terms of the dissociation thesis. Automatisation occurs when a cognitive or motor task that initially places heavy demands on the cogniser, and is relatively slow and effortful, progressively becomes fast, effortless and "unconscious". Standard accounts treat automatisation as the gradual withdrawal of attention, where attention is understood as a limited capacity resource closely associated with consciousness. An implicit assumption here is that the very same representations and processes feature in the unconscious, skilled performance of a task as are involved when it is controlled consciously. This approach has been reasonably successful in accounting for the observed differences between automatic and non-automatic performance, although some researchers have questioned the view that automatic processing is free of attentional limitations (see Logan 1988, pp.492-3 for discussion). It's not my intention to critique the standard account here. Rather, I want to discuss an alternative approach to automatisation; one that is compatible both with connectionism and with a denial of the dissociation thesis.

To account for automatisation without invoking the dissociation thesis it is necessary to establish that automaticity can be satisfactorily explained in terms of unconscious processes that are free of explicit representations. This means rejecting the claim that what occurs during the automatic performance of a task is computationally similar to what goes on during its non-automatic execution. Logan (1988, 1990) countenances such a rejection in his *instance* theory, which treats automatisation as the replacement of algorithm-based processing with memory retrieval. On this account, performance becomes automatic when it involves "single-step direct-access retrieval of past solutions from memory" (1988, p.493). Non-automatic performance, on the other hand, reflects reliance on a general algorithm. Thus, Logan's account contrasts with traditional approaches "which assume that the same algorithm is used throughout practice; processing remains the same but uses less attention…or is done in parallel rather than in series" (1990, p.4). Instead, Logan suggests that automatisation involves a transition from algorithm-based processing to memory-based processing, and "reflects the development of a domain-specific knowledge base; non-automatic performance is limited by a lack of knowledge rather than by the scarcity of resources" (1988, p.501).

A simple example of automatisation that supports Logan's theory is the development of arithmetic skills in children. Initially children add by counting. They employ a (tacit) algorithm such as: (1) count the first addend with, say, marbles; (2) count the second addend using more marbles; and (3) combine the marbles and count the collection. This process is slow, and relatively difficult. Later on, addition with single digit numbers becomes a quick and easy act of recall. It seems unlikely that expertise with addition relies on unconscious counting. Similarly, typing starts as a painfully slow process of visual search for the appropriate keys. It then goes through two (or more) stages of refinement. First, there is a growing familiarity with the placement of the letters on the keyboard, such that visual search becomes progressively quicker. Second, as the typist moves on to "touch typing", visual search is dispensed with entirely, and there develops a fairly direct connection (at least phenomenologically) between recognising words or word-fragments (e.g., 'tion', 'ing', etc.), and apprehending the "shape" of the required motor responses. The first of these refinements is pretty clearly a case of memorising the required letter-to-keyboard correspondences (although, I suppose the hypothesis that it involves rapid, unconscious

search is not completely ruled out by first-person evidence). To treat the second refinement as a memory-based process is, perhaps, more controversial – one might plausibly maintain that it relies on a rapid, unconscious sequence of instructions to enact individual finger movements. But here quantitative studies are relevant. And it appears that Logan's theory fits the data as well as, or better than, traditional models for a range of learning tasks.[121] Specifically, it accounts for the widely observed power-function speed-up in automatisation, and in so doing predicts a constraint between the mean and standard deviation of reaction time that was observed in a number of experiments. More significantly, Logan was able to show that a repetition effect is specific to individual stimuli, as the instance theory predicts. Transfer to novel stimuli was found to be poor, a result that creates difficulties for standard accounts of automatisation.

The beauty of Logan's theory of automatisation is that it's congenial to both connectionism *and* the connectionist vehicle theory of consciousness. Connectionism can accommodate the progression from relatively slow and effortful non-automatic performance, to fast, effortless, automatic performance, in terms of a gradual move from sequential multi-network processing, to processing within a single specialised network. The idea is that novel tasks – both motor and cognitive – are initially handled by co-operating sets of less specialised networks, which proceed algorithm-style to task completion, via a sequence of network stabilisations. Repeated operation of such multi-network solutions, in conjunction with innate learning mechanisms, leads to the training of a specialist network (or networks) that can enact the same tasks in a single-step fashion (Smolensky makes essentially this suggestion in his 1988, p.12). On this account, the gradual change in performance seen during automatisation reflects a growing reliance on single-network pattern-matching (which corresponds to what Logan calls "single-step direct-access retrieval of past solutions from memory") and a tendency to abandon more computationally expensive multi-network solutions. While highly conjectural, this account of automatisation is very natural from a connectionist perspective, especially in view of the fact that complex tasks such as text-to-phoneme transformation, facial recognition, and past-tense production, are known to be achieveable by individual networks.

Logan's account of automatisation, and its proposed connectionist implementation, comport well with the phenomenological evidence. Automatic processes are effortless because they involve single-step retrieval of solutions from memory. By contrast, non-automatic processes are relatively effortful because they require a sequence of conscious steps in order to arrive at the same result. Automatic processes are unconscious, because they are retrieval processes, whereas non-automatic processes involve numerous conscious intermediaries that correspond to the steps (and sometimes even the rules) of whatever algorithm they realise. Having said this, it is important to be clear about what it means to say that automatic processes are "unconscious". There is actually a good deal of phenomenology connected with skilled behaviours.[122] In a simple automatic process, such as adding two numbers, one must clearly be conscious of the input (the addends) and the output (the sum) in order to succeed at the task. What is unconscious is the retrieval mechanism that links addends to sum. This mechanism bridges the "cognitive gap" (as Lloyd (1991) calls it) between conscious start and end states. More complex automatic processes, such as skilled typing, depend on long trains of conscious stimuli and responses,

---

[121] Logan discusses numerical studies of a lexical decision task, and "alphabet arithmetic" in his 1988.

[122] The flip-side of this is that there is also a lot of *un*conscious processing connected with consciously enacting an algorithm of some kind. See below for more on this point.

without which things would rapidly grind to a halt.[123] Here, it is each of the individual stimulus/response pairs that is linked by an unconscious retrieval mechanism. During the long periods of laborious repetition involved in learning such a skill one's experience of the task undoubtedly changes. But consciousness does not drop out, rather, it is transformed: one is no longer conscious of the individual muscle movements involved; one experiences the whole process in a more coarse-grained fashion, in which "chunks" of input produce "chunks" of output.

All of this still leaves me facing the dissociation thesis, because, even assuming that Logan's account of automatisation is the right one, it is an open question whether retrieval of past solutions from memory involves unconscious explicit representations. However, notice that the connectionist implementation of Logan's theory I sketched above does in fact allow one to dispense with such states. On the connectionist model, a task that has been automated is one that has been transferred to a single, specialised network. Such a network receives as input a stable set of signals over its input lines (corresponding to some consciously apprehended stimulus condition), and produces in response a stable activation pattern. In so doing it effects a retrieval from memory of an appropriate solution to the input condition. The only explicit information implicated in this process is connected with perceptual input and motor/cognitive response, both of which are phenomenally conscious. Memory retrieval itself, which is clearly unconscious, doesn't involve explicit intermediaries, because it consists in the relaxation process that *generates* an explicit response. This account proposes nothing in the way of explicit representation that is not reflected in our actual experience of skilled behaviour. Whenever explicit representations are posited (the inputs to an automatic process, and the generated responses) we discover corresponding elements of conscious experience. And whenever unconscious processes enter the story, these may plausibly be construed as network stabilisations, which processes implicate no explicitly represented information. Automatisation begins to look like a process whereby a range of explicit states connected with slow, effortful, algorithm-style thinking, gradually drop out, to be replaced by rapid, effortless relaxation processes in neurally realised PDP networks. From this perspective the standard account of automatisation simply begs the question against an independently plausible approach to the phenomenon.

Perceptual Processing Much of Section 5.3 was devoted to a discussion of the evidence for the dissociation thesis supposedly provided by contrastive studies of perceptual processing (studies of dichotic listening, visual masking and blindsight). I attempted there to throw some doubt on the usual interpretation of these studies. But even if we regard the contrastive analyses as inconclusive, there is a more general issue as to how a connectionist vehicle theorist might account for perception. The phenomenon of perception – the production of coherent, unambiguous experiences of physical objects, given impoverished, often ambiguous stimuli – leads most theorists directly to an account rich in both unconscious rules, and unconscious representations.[124] Even so, I believe a connectionist can aspire to undermine the influence of the dissociation thesis in this domain, without abandoning the powerful insights of the cognitive tradition.

---

[123] This gloss on complex automatic processes is not intended to exclude the possibility that such processes may, in some cases, be largely internalised, and therefore not dependent on constant feedback from the environment. See my discussion of calculation, below.

[124] Prominent pieces in this tradition are Chomsky 1980, Fodor 1983, Marr 1982, and the literature of the "problem-solving" approach to perception exemplified in the work of Rock (1983).

The essential concern for a connectionist vehicle theorist is that the available processing resources are insufficient to account for perception. However, there are two reasons to regard this worry as overblown.

First, it is not clear that a connectionist needs anything like the number of explicit representations employed in traditional models of perception. Consider again the behaviour of NETtalk (see Chapter 2 for a fuller discussion). 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 pretty minimal network, by brain standards. But when PDP techniques are applied to large-scale perceptual systems (such as the visual system[125]), 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 data in a monolithic, reflex-like fashion.[126]

Second, my analysis in Chapter 3 suggests that instantaneous 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. Consequently, a connectionist vehicle theorist actually has a great deal of explicit representation to play with when framing theories of perception. And indeed, while PDP models of, say, vision, eschew explicit rules, it seems likely that they will be committed to the existence of a hierarchy of explicit data structures, as classical models are.[127] No doubt some will count this as a mark against the connectionist vehicle theory of consciousness, since it is orthodox to assume that the data structures generated in early perceptual processing don't directly contribute to phenomenal experience. But one wonders why anyone would assume this. Zeki 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). And it strikes me as reasonable to suppose that the contents of those data structures 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 supports this view. With regard to the "lesser sensations" of vision, the low-level visual representation of edges, figural boundaries etc., he claims:

---

[125] See, for example, Arbib & Hanson 1987, Lehky & Sejnowski 1990, Marr & Poggio 1976.

[126] One respect in which NETtalk is not paradigmatic, so far as perceptual processing is concerned, is that it is a strictly feedforward network. Such networks are unable to resolve ambiguities in incoming signals, or arrive at alternative interpretations of a single stimulus. However, recurrent networks can deal with these tasks. See Churchland 1995, pp.107-14, for discussion.

[127] Lloyd (1991) points this out. He notes that none of the current PDP models of vision simulates the entire visual process:

> They suggest, in fact, a confirmation of the standard computational model of vision, a hierarchical progression from retinal patterns through what Marr (1982) called the "primal sketch", assembled of edges and gradients, on up to a three-dimensional model of a world of meaningful objects. At each stage in the hierarchy we find representations, and the various [PDP] networks…simulate isolated transitions in the hierarchy, rather than the whole process. (p.454)

> 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.

These comments by no means do justice to the wealth of phenomena that has led theorists to rely heavily on explicit, but unconscious representations in their models of perception. In this connection I mention the processing of ambiguous stimuli, and studies of visual illusions. The former is nicely illustrated by our capacity to arrive at a single consistent reading of a sentence containing one or more ambiguous words. Typically only interpretations that cohere with the local semantic context ever occur to us consciously. It is customary to adopt a "hypothesis-testing" model of this phenomenon, in which various interpretations are tried out unconsciously, and only one which satisfactorily resolves all linguistic ambiguities is forwarded for conscious appreciation. However, connectionism offers another approach to the resolution of perceptual ambiguities. In PDP models of such phenomena, rather than supposing that alternative hypotheses must be explicitly formulated and tested for consistency, the final coherent interpretation is arrived at via a process of "relaxation search". This involves activation passing among a set of units with multiple feedback connections, such that stable activation is not achieved until a best-fit solution to the input condition is achieved. Without considering a detailed example[128], the essential point here is that, until the relevant networks stabilise, no explicit representation is generated by this kind of process. If such an account can be made to stick then there's no need to suppose that perceptual disambiguation involves the explicit but unconscious representation of information.

The study of perceptual illusions teaches us a great deal about the brain, primarily because it reveals the kinds of "assumptions" that are built into our perceptual systems. For example, when one views gray-scale images that incorporate a shading gradient, objects which are lighter at the top typically appear to be convex, while those which are lighter at the bottom appear to be concave. This is the correct way to interpret three dimensional objects (i.e., objects which are genuinely concave or convex) that are illuminated from above. The alternative "reading" of such an image is difficult to achieve. Thus, it appears that a default assumption is built into the visual system to the effect that light sources are normally overhead (Ramachandran 1988). One might think that such an (unconscious) assumption must be explicitly represented somewhere in the brain, and so creates a problem for the connectionist vehicle theory of consciousness. It is well known, however, that operations properly described as following a rule, or instantiating a perceptual convention of some kind, need not be realised by a computational mechanism which explicitly encodes that rule or convention. The lesson of NETtalk is very general: *rule-describable* behaviour can arise from a system that is not *rule-governed* in any strict sense.[129] The assumptions built into our perceptual systems, in particular, may be realised in their "wiring", i.e., in the patterns of connectivity that determine how those systems operate. Such patterns constrain our perceptual apparatus to afford us with certain interpretations of the input, and not others,

---

[128] But see, for example, the interactive activation model of word recognition discussed in McClelland, Rumelhart & Hinton 1986, pp.20-3.

[129] See Devitt & Sterelny 1987, Ch.8 for further discussion of this point.

without realising perceptual conventions explicitly. Consequently, the study of perceptual illusions doesn't provide us with strong grounds for rejecting the connectionist vehicle theory of consciousness.

Reasoning In Chapter 3 I examined the nature of "higher-thought": thought processes that are largely internally driven, and involve extended sequences of conscious states. The principal kinds of higher-thought are nicely captured in Johnson-Laird's (1988) taxonomy of thinking, which distinguishes among *association* (which has no goal), goal-driven activities like *calculation* (deterministic), *deduction* and *induction* (both non-deterministic), and *creativity*. (See Section 3.2 for further discussion.) These forms of higher-thought feature in both day-to-day problem solving and in formal reasoning. Not all higher-thought is symbolic, in the sense of being conducted in a natural language (inductive reasoning, for example, often seems to involve the use of analogies that resist expression in symbolic terms), but reasoning processes that employ a symbol-system of some kind are a very significant subclass of higher-thought. Thus, when framed in linguistic or mathematico-logical terms, higher-thought tends to involve extended sequences of conscious symbol-manipulations. Moreover, such symbol manipulation often conforms to rules, which may themselves be conscious (e.g., "i before e except after c").

It appears, then, that higher-thought sometimes takes the form of serial, rule-governed symbol processing. Smolensky characterises such thought as the province of the *conscious rule interpreter*: a virtual machine implemented in PDP hardware which can be idealised as universal (capable of executing any effective procedure) (1988, p.4). He distinguishes this machine from the *intuitive processor*, the virtual machine responsible for all behaviour that isn't the outcome of conscious rule application. The intuitive processor, he suggests, governs "practically all skilled performance", which includes "[p]erception, practiced motor behavior, fluent linguistic behavior, [and] intuition in problem solving and game playing" (ibid., p.5). It is standard practice in cognitive science to assume that the phenomena governed by these two machines can only be fully accounted for by modelling the unconscious on the conscious rule interpreter, i.e., by populating it with a great many explicit representations. The present section (7.3) is a connectionist challenge to this assumption. I have tried to show that, insofar as perception and thought are associated with the explicit representation of information, such information may plausibly be aligned with the contents of consciousness.

One might object to this project on the grounds that, while certain aspects of perception and skilled behaviour may be explicable without recourse to unconscious, explicitly represented information, higher-thought is far too representation hungry to permit a denial of the dissociation thesis. In particular, it is unreasonable to expect that those explicit representations with *conscious* contents could possibly suffice to account for the full range of our cognitive capacities. But we can undermine this objection in two ways.

First, as I intimated above in my discussion of perceptual processing, moment by moment consciousness is thick with distinct contents. Not only is it a multi-modal aggregate of perceptual experiences, but it also simultaneously incorporates a great many levels and kinds of understanding experience (see Section 3.2). On the uncontroversial assumption that all of this experience somehow depends on the vehicles of explicit representation in the brain, it appears that consciousness actually provides us with a good deal of representation with which to account for human cognition. The felt need for an extremely rich cognitive unconscious is, to some extent, a consequence of downplaying the significance of the contents of consciousness.

Second: it may be that cognition need not involve as many explicit representations as has been traditionally assumed. Clark (1997) has recently argued that in the performance of

many tasks humans reduce their computational load by relying on clever mind-world interactions. We not only use a great many external memory devices, hence reducing the amount of information that must be carried on board, but we also "actively structure and operate upon our environment so as to simplify our problem-solving tasks" (p.67). To give a simple example: in the game of scrabble, players tend to arrange and rearrange their pieces into significant parts of words. This restructuring of the linguistic environment actively promotes word recall (especially if our basic cognitive resource is a set of content-addressable pattern-completing machines), and significantly reduces cognitive load compared to the same task undertaken with a random ordering of input. (ibid., p.64) The upshot is that human cognition probably involves far fewer explicit representations than theorists have generally supposed.

More needs to be said here. In order to respond convincingly to the objection above I need to make some specific remarks about intuition in problem solving, and about the role of explicit representation in the operations of the conscious rule interpreter.

To begin with, there can be no hard and fast distinction, in my view, between intuitive problem solving and conscious rule-following. Every case of conscious rule interpreting involves the intuitive deployment of appropriate rules. By this I mean that even in a very rigid thought process (such as calculation) there is some point at which skilled, context sensitive retrieval of information is involved.[130] Since such retrieval is unconscious it is clear that both the conscious rule interpreter and the intuitive processor rely on rapid, unconscious processes of some kind. Intuition thus enters into all forms of problem solving. What are we to make of such intuition? What, in particular, are we to make of the very considerable range of cases where a skilled practitioner can simply "see" what to do next (which move to make, which rule of inference to apply, what diagnosis to make), or where long and laborious consideration of some complex problem terminates in sudden insight (I'm thinking here of the kind of creative thinking that typically incorporates incubation periods – see Section 5.2). The two kinds of processes are distinct – the one being rapid, and fairly reliable, the other, time-consuming and not guaranteed to succeed – but both have inclined many theorists towards an unconscious rich in explicit representations. Connectionism suggests that we model the former as a process of relaxation search, which begins with the perceived problem situation, and ends with a stable pattern of activation that constitutes its solution. This settling process may involve multiple networks, but need not generate any explicit representations intermediate between input and output. The networks involved in this rapid, intuitive aspect of problem-solving may be conceived as the product of a learning mechanism like the one I sketched in my discussion of automatisation. If such connectionist models can be made to work (and the jury is still out on this one) then such forms of problem-solving present no special problem for the connectionist vehicle theory of consciousness.

Promising work in this direction is reported by Bechtel and Abrahamsen (1991, pp.163-74), who designed localist PDP networks to assess arguments in propositional logic, and supply missing premises in enthymemes, respectively. Student performance on the first of these tasks, which involves identifying the form of an argument, and deciding whether or not it is valid, was found by Bechtel to improve only slowly. Early homework contained many errors. Results improved with correction and criticism, although most students did not achieve flawless performance, some still making up to 25% of judgements in error. Network performance, although unrealistic with regard to the amount of training required, showed a similar pattern, thus supporting the contention that "human competency in formal

---

[130] Smolensky proposes that "in a subsymbolic rule interpreter, the process of rule selection is intuitive" (1988, p.13).

reasoning might be based on processes of pattern recognition and completion that are learned gradually as by a network" (ibid. p.173).

A further reason to take seriously the idea that skilled intuitive judgements actually consist in processes of pattern matching, or relaxation search, is that the alternative – rule-governed manipulation of recursive symbol structures – gives rise to the notorious difficulties associated with the frame problem. We really have no idea how to get a classical system to rapidly update and retrieve knowledge in a globally-sensitive fashion, given classicism's commitment to universal causal discreteness (Section 2.3). All the current indications are that too much explicit representation positively gets in the way of fast, flexible, contextually-nuanced performance. PDP systems, on the other hand, process information in a globally holistic manner, by virtue of the fact that the processing substrate of a PDP network is identical to the substrate responsible for encoding its store of knowledge.[131] There is thus a significant sense in which everything that a PDP system knows is simultaneously brought to bear to determine processing outcomes, *without* each potentially significant item of information having to be explicitly retrieved and processed. If we suppose that intuitive reasoning is implemented in PDP architecture then the global sensitivity of such processes doesn't appear quite so mysterious. At the same time we see that connectionism's potential to account for this feature of thought, if it does not actually demand rejection of the dissociation thesis, is at least complemented by that proposal. For it is in the unconscious that global sensitivity must ultimately be worked out, strongly suggesting that we move towards a model of unconscious processing in which the role of explicitly represented information is kept to a minimum, or better, entirely eliminated. Denial of the dissociation thesis thus goes hand in hand with our best current shot at solving the frame problem.

Let us turn now to the kinds of intuitive processes involved in creative problem-solving: non-deterministic reasoning that commences with a reasonably well defined goal (such as, *find the damn house-keys*, or *prove Fermat's last theorem*), but has no clear-cut starting point. One important characteristic of this form of reasoning is that it's a relatively long-winded process, by virtue of incorporating numerous distinct conscious episodes. These need not be temporally contiguous to ensure success (relevant thoughts can occur at all sorts of odd and scattered times, especially when trying to sleep at night!). Creative problem-solving is also distinctive for its recognisable phases, which Wallas (1931) termed: *preparation*, *incubation*, *illumination*, and *verification* (see Section 5.2 for brief descriptions). It is during the preparation phase that one first "loads" a problem into one's head, using whatever means are appropriate to the task at hand (be they symbolic, or non-symbolic, formal, or informal), and then makes preliminary efforts to solve it. The latter involves attempting to establish some connection between the goal-state and existing knowledge. One may proceed by analogy with a similar problem, or may attempt to find a way of formulating the problem that renders it more tractable. It's well known that adopting the right formalism will often significantly reduce the difficulty posed by a problem, but clearly it is the process of discovering an appropriate problem "representation" that constitutes the real creative challenge.[132] Preparation is followed by incubation, and very often by further cycles of preparation and incubation, before the crucial insight emerges. Incubation is the phase during which a problem is set aside, and the problem-solver rests, or turns his attention to

---

[131] And this implies that there's no real distinction between potentially explicit and tacit representation in a PDP system, as I pointed out in Chapter 2.

[132] Any good text-book in AI or cognitive psychology will provide numerous examples of problems whose solution depends on discovering an appropriate representation (i.e., formulation) of the task. See, for example, Rich & Knight 1991, or Anderson 1995.

some other activity. There is a great deal of anecdotal evidence, and some experimental support, for the utility of this phase in problem-solving. Illumination – the spontaneous emergence of some key insight – often follows directly upon the heels of incubation, although it is fair to say that not every problem is solved in this manner. Even though one may feel quite certain at this point that the problem is solved, it is usually necessary to verify one's idea, because sudden "insight" sometimes turns out to be bogus.

What kind of story can one tell about creative problem-solving consistent with a denial of the dissociation thesis? I must admit at the outset that I have only the crudest of accounts to offer. To begin with, it is clearly necessary to reject the view that what goes on during incubation periods involves unconscious problem-solving. But I've already argued for this, on the grounds that there are numerous plausible explanations of the incubation effect which don't involve treating incubation as "unconscious work", and at least one of which is supported in the lab (Section 5.3). The key to the account concerns the relationship between conscious and unconscious activity. During the preparation phase of problem-solving there is a great deal of intuitive processing going on. We may think of this as an iterated series of single-step operations, leading from one conscious state to the next.[133] Each such "operation" is, of course, enormously complicated – consisting of relaxation processes involving an enormous number of neurons – but for all that it may fail to implicate any unconscious explicit representations, if the account of skilled intuition I urged above has any merit. According to the connectionist vehicle theory of consciousness, the conscious states implicated in problem-solving are explicit representations: stable patterns of activation in neurally-realised PDP networks. They are the *product* of unconscious processing, which is determined by the connections and connection weights that support the inexplicit storage of information in the brain. Equally, however, the unconscious is affected by conscious activity. Connection weights are labile – they are modified in order to store new information – and on anyone's story a principal mechanism of such modification is going to be the presence of explicit representations. Consequently, conscious activity *itself* modifies the very substrate that determines the possible contents of experience. When you attempt to solve a complex problem you are thus embarking on a journey through weight space. Its intended destination is a set of networks with the capacity to generate the problem's solution. Each conscious state is the product of current activity and input, and the unconscious background of potentially explicit information. The vital part played by these conscious episodes is the resetting of connection weights, so whenever you attend to your problem the substrate that supports unconscious processing is modified. Subsequent attention will find the network dispositions subtly altered, such that you may be closer to a solution. With a little luck the problem-solver's journey will end at the right point in weight space. It just remains to think a triggering thought, and cry Eureka!

Finally, I will say a few words about the nature of the conscious rule interpreter. It is important to recognise that the expression "conscious rule interpreter" is ambiguous between thought that involves the conscious deployment of rules, and conscious thought that merely conforms to syntactic rules of some kind. In the latter case the cogniser need not be conscious of the relevant rules in order for them to accurately capture the process in question, she must simply pass through a sequence of conscious steps prescribed by those rules. Both kinds of thought are significant, and both are a potential embarrassment for a connectionist vehicle theory of consciousness. I particularly have in mind here the kind of thinking involved in performing mental arithmetic, or carrying out a formal proof, in which the cogniser must perform operations that are sensitive to the formal structure of consciously

[133] These are not "total" conscious states, since there will likely be a great many elements of personal consciousness that are not relevant to the problem under consideration (e.g., proprioceptive sensations, background noises, the colour of one's coffee mug, etc.).

apprehended symbols.[134] The problem is not that such thought looks like serial, von-Neumanesque symbol-processing. That would only upset a connectionist of an extremely radical disposition. It is, rather, that rule-governed conscious thought may turn out to depend on processes that implicate *unconscious* explicit representations. And that would clearly be deadly for a connectionist vehicle theory of consciousness. But I don't think we need be too bothered by this possibility, because connectionists have developed a plausible hypothesis about conscious symbol-processing that should still the fears of both the radical connectionist and the connectionist vehicle theorist. The basic move is to go "externalist" about the conscious rule interpreter, an idea first mooted by Rumelhart, Smolensky, McClelland and Hinton (1986).[135]

Consider the process of pen and paper arithmetic. One manipulates a set of external symbols according to a set of simple, recursive rules (e.g., write the multiplicands one above the other, multiply the rightmost digits of the two numbers, write down the first digit of the product and carry any further digits, and so on). But in so doing one need not represent those rules, either consciously or unconsciously.[136] Instead the process is very much driven by environmental input, and relies on the basic pattern matching resources of PDP systems. Having written down the two numbers one recognises what move to make first (via a single-step recognition process of some kind), and then carries out that move using instance based retrieval (i.e., matching the pair of single digit multiplicands with their product). The result, once inscribed on paper, then provides the stimulus for the next step in the sequence, and so on. It is important to recognise that this process relies not only on simple pattern-matching resources, such as those involved in associating summands with their sum, but on a hierarchy of interacting automatisms, including some that respond to complex environmental conditions by directing visual attention, thereby eliciting lower-order automatisms, and so forth.[137] Calculation thus comes out looking like a fancy, hierarchically-chained set of automatisms realised in PDP architecture. Formal deductive and inductive reasoning can hopefully also be made out in this way, although such processes, being non-deterministic, require PDP mechanisms that provide a one-to-many mapping between input and output. A connectionist vehicle theorist can take comfort in this model, since the only explicit representations it seems to require are those connected with consciously apprehended symbols and symbol arrays. Automatisms connecting perceived environmental conditions of various kinds with appropriate responses can be treated as simple pattern matching exercises, or as settling processes in recurrent networks, and as such don't depend on unconscious explicit states.

There is more to say here, because it is, of course, possible to do pen-and-paper arithmetic in one's head. The externalist move then loses its force. However, Rumelhart et al.

---

[134] Recall that a digital computer sometimes deploys explicit rules, and sometimes acts in accordance with rules that are merely tacit. In both cases the resulting behaviour counts as symbol-processing insofar as the computational operations involved are driven, at some level, by the syntactic structure of the machine's representations. The point is that conscious thought, when it strictly conforms to syntactic rules of some kind, counts as symbol-processing for the same reason, irrespective of whether it involves explicit (hence, conscious) rules, or rules that are merely tacit (and therefore unconscious).

[135] Discussion and elaboration of this idea can be found in Smolensky 1988, Clark 1989, Ch.7, and Bechtel & Abrahamsen 1991, Ch.7. The term "externalism" is due to Bechtel and Abrahamsen.

[136] A novice may in fact deploy a set of conscious rules when doing arithmetic, but with increasing expertise these tend to drop out.

[137] Such mechanisms are presumably the product of a long process of learning, which itself depends on earlier hierarchies of symbol-driven automatisms, including especially those involved in the intuitive deployment of language. See Smolensky 1988 for more on this theme.

have another clever move up their sleaves. They suggest (after the fashion of Vygotsky 1962) that when one performs mental arithmetic, rather than sending output, as it were, direct to the motor effectors, one *internalises* the process by introducing a model of the external symbolic environment. Instead of storing the result of a multiplication, say, using some external medium, one places it in working memory, rehearsing it as necessary to provide an internal stimulus for the next stage of calculation. This is real, sequential symbol-processing, and Clark may be right to suggest that a correct, and in some sense complete, psychological account of such tasks can be given in classical terms (1989, p.136). In some circumstances our brains appear to be simulating a conventional (digital) architecture, and it may therefore be necessary to adopt a *hybrid* model of human cognition. Note, however, that the digital virtual machine that we perhaps realise when doing mental arithmetic does not, according to the model of Rumelhart et al., rely on any deeper level of symbol-processing – it is not built with digital hardware. Rather, it depends to a very significant degree on intuitive processing, by virtue of being directly realised in PDP hardware. Moreover, it is computationally expensive symbol-processing, relying as it does on both visual and auditory resources (the visuo-spatial and articulatory loops of working memory), in addition to an array of specialised task-specific networks. And if my ruminations in this section are borne out, it may be that this is the only form of symbol-processing our brains ever engage in (a sentiment echoed by Rumelhart et al. (1986, p.46)). If this is the kind of hybrid model that the evidence requires of us then it's not one that should frighten even a radical connectionist, for it banishes symbol-processing from the underlying computational architecture, relegating it instead to "ingenious icing on the computational cake" (Clark 1989, p.135). Neither should such a model be of concern to an advocate of the connectionist vehicle theory of consciousness, because what has emerged here is that the conscious rule interpreter gives rise to explicit representations only to the extent that it generates a sequence of rule-governed conscious states. There are no unconscious symbols, or explicit representations, of any kind.

The Connectionist Unconscious

In Chapter 5 I remarked that on the standard (classical) model of the unconscious, cognition is subject to a kind of representational and processing homogeneity: explicit representations feature as legitimate vehicles of both conscious and unconscious contents (the representational homogeneity); and unconscious processes are understood by analogy with the causally discrete, structure sensitive operations we observe in much higher-thought (the processing homogeneity). A quite distinct picture of the unconscious has emerged in this section, in line with my attempt to show that connectionism will tolerate a denial of the dissociation thesis. Unlike its classical counterpart, the connectionist unconscious I've sketched is devoid of explicitly represented information. So connectionism breaks the representational homogeneity of classicism. Explicit representations are no longer treated as legitimate vehicles of both conscious and unconscious contents. Only *conscious* information is explicit; unconscious information is all and only *inexplicit*.

As for unconscious processing, again the homogeneity of classicism is banished. Conscious and unconscious processes are no longer treated on the same model. While conscious processing involves the *inter*-network flux of explicit information, and thus takes on a seriatim aspect, unconscious processing involves the causally holistic *intra*-network operation of potentially explicit information. Unconscious processing is therefore responsible for the *generation* of explicit vehicles of content, not the *manipulation* of such representations.

7.4 Informational Access

I will finish this chapter with a few brief remarks about the relationship between inter-network information processing relations and phenomenal experience, which will help to further illuminate the connectionist vehicle theory of consciousness.

Process theorists seek to explain phenomenal consciousness in terms of computational activities that privilege certain representational vehicles over others (see Chapter 6). What distinguishes explicit representations whose contents are conscious from those whose contents are not, according to the process theorist, is that the former – but not the latter – have some special computational role, or are subject to specific kinds of computational processes. The emphasis is on what explicit representations *do*, rather than on what they *are*. One natural way to develop this idea, given the seemingly intimate relationship between phenomenal experience and the rational control of speech and action, is to suppose that consciousness is somehow tied up with the transfer of information in the brain. A number of prominent theorists have been tempted to suppose, in particular, that consciousness is constituted by *rich* information processing relations within the brain. I call this the "access thesis". It is the claim that those mental contents are conscious whose vehicles have *rich* and *widespread* informational access to a subject's cognitive economy. Both Baars (1988) and Dennett (1993, 1996) seem to be committed to this kind of view.[138]

Assuming my account, the access thesis gets things exactly backwards: there is no path leading from rich informational access to consciousness; consciousness precedes, and is responsible for, such access. To see this, consider the basic claim of the connectionist vehicle theory of consciousness: that phenomenal experience is identical with explicit representations – i.e. stable patterns of activation – in PDP networks realised in the head. Information encoded explicitly in a PDP network is conscious in virtue of a high-level physical property of that network, not in virtue of any rich access relations it might enjoy. This is a direct denial of the access thesis. Consciousness is an intrinsic, physical *intra*-network property, on my account. What, then, is the connection between consciousness and informational access? Access relations inhere in larger systems; collections of interconnected networks, and depend on the capacity of each network to have effects on the others. Stability of activation plays a pivotal role in such relations. A recurrent network processes information via a mechanism known as "interactive search", which goes to completion when the network "relaxes" into a stable pattern of activation in response to a stable array of inputs over its input lines (see Chapter 2). Processing is not complete until stable activation has been achieved, thus, stability is central to the connectionist account of cognition (especially on the plausible assumption that interactive network architectures are dominant in the brain). Moreover, connectionism suggests that stable activation has a material role to play in the information processing relations that obtain between distinct neural networks in the brain: one network is far more likely to have significant effects on another (i.e., play a determining role on its course of processing) if its state of activation has stabilised. As I

---

[138] Unfortunately, when talk of rich informational access is in the air some read this as reference to what Block calls "access-consciousness" (1995). It is vital not to confuse access-consciousness with the access thesis. The former is a *sort* or *kind* of consciousness. Here's how Block characterises it:

> A state is access-conscious (A-conscious) if, in virtue of one's having that state, a representation of its content is (1) inferentially promiscuous (Stich, 1978), i.e. poised to be used as a premise in reasoning, and (2) poised for rational control of action and (3) poised for rational control of speech. (1995, p.231)

Thus, a content is A-conscious if its representational vehicle has rich and widespread informational access in a cognitive system. The latter is *not* a special kind of consciousness – it is a thesis about how phenomenal experience *is to be explained*. In particular, it is the suggestion that a content is phenomenally conscious *when* its representational vehicle has widespread informational access.

remarked in Section 7.1, stability tends to beget stability. The importance of this, in the present context, is that phenomenal experience is *constituted* by stable patterns of activation in neurally-realised PDP networks, according to the connectionist vehicle theory of consciousness. Consequently, the capacity of a network to have significant effects in the wider cognitive system is dependent on explicit tokenings. And this implies that *informational access depends on phenomenal experience, not the reverse*.

It is vital to realise, however, that the presence of phenomenal experience is necessary, but not sufficient, for inter-network access. Explicit tokenings are not guaranteed to have informational 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).[139] Thus while consciousness facilitates such information processing relations, it can exist in their absence. One can be conscious of some content, yet unable to act on the information. This is perhaps the situation of a drunk driver who loses control of his car. In this case the necessary information, although part of conscious experience, simply doesn't get through to the relevant control mechanisms. The reverse typically doesn't obtain, however, since rich and widespread access can't exist without some associated phenomenology.[140] Thus, Sacks' disembodied lady (see Section 4.1), who lacks all normal body phenomenology, suffered a complete breakdown of bodily control. Without body awareness (even if only in the periphery of attention), those control mechanisms responsible for the maintenance of body posture and balance can't get access to the information they need. The posture she eventually did achieve was the result of painstakingly learning to use visual feedback to provide the necessary information. This required the formation of new connections between the visual and motor systems, so that existing stable activation (and hence phenomenology) in the visual cortex could be used for motor control (Sacks 1985, pp.42-52). To repeat my theme: phenomenal experience is necessary but not sufficient for informational access; an appropriate set of inter-network connections being a further requirement.

This intimate relationship between informational access and phenomenal experience suggests an explanation for the "fallacy" identified by Block (1995): the tendency to transfer functions of A-consciousness (rich informational access) to P-consciousness (phenomenal experience). At the same time it explains the popularity of the access thesis. Regarding the latter: since, on the connectionist vehicle theory of consciousness, rich access relations are always associated with phenomenology, it is tempting to suppose that it is rich and widespread informational access that constitutes consciousness. But this is to put the cart before the horse. Phenomenology enables access, not the reverse. As for Block's fallacy: failures of access will often (but not invariably) be connected with loss of phenomenal consciousness, given my hypothesis, so it is easy to find oneself "jumping from the premise that 'consciousness' is missing - without being clear about which kind of consciousness is missing - to the conclusion that P-consciousness has a certain function" (1995, p.27). From the perspective of the connectionist vehicle theory of consciousness it is very tempting to transfer functions of access (A-consciousness) to phenomenal experience, given the close connection between the two. Indeed, on that account stable activation is the modus operandi of access, so lack of phenomenal consciousness of some content precludes the realisation of

---

[139] The same is true, mutatis mutandis, of classical systems, although there phenomenal experience can't line up with the explicit representation of information. As a consequence, a classicist who doesn't subscribe to the access thesis can presumably allow the possibility of rich informational access *without* any associated phenonemology. On the connectionist vehicle theory of consciousness this is not possible.

[140] Notwithstanding the effect of truncating intra-network processing just prior to the production of a stable pattern of activation, as in visual masking (see Section 7.3 above). The effects in this case are neither widespread nor particularly significant.

access functions in which it might be involved. Loss of consciousness often *explains* loss of access. Block identifies this kind of position when delineating the various possible relationships between A- and P-consciousness. But it doesn't fall foul of the fallacy he warns against:

> Of course, it could be that the lack of P-consciousness is itself responsible for the lack of A-consciousness. If that is the argument...I do not say "fallacy". The idea that P-consciousness is responsible for the lack of A-consciousness is a bold hypothesis, not a fallacy. (1995, p.242)

The hypothesis is "bold", I suggest, because, as Block remarks: "there is some reason to ascribe the opposite view to the field as a whole" (1995, p.242). And it is not fallacious, because it doesn't involve any confusion between access- and phenomenal-consciousness.

The connectionist vehicle theory of consciousness holds, therefore, that phenomenal consciousness facilitates informational access relations, but can exist in the absence of such relations. A classicist could subscribe to this position if she treated consciousness as some kind of functional gateway to mechanisms of access, but that would amount to a *further* hypothesis about the functional architecture of the classical system. For a connectionist vehicle theorist this relationship between informational access and consciousness emerges in a natural and *principled* way from a hypothesis concerning the representational status of phenomenal experience.[141] Once one assumes the identity between explicit representation and phenomenal experience it is unavoidable that rich access relations should be contingent on consciousness. To borrow a metaphor from Block: "P-consciousness is like the liquid in a hydraulic computer, the means by which [access] operates" (1995, p.242).

[141] One might object here that on my account potentially explicit information, which *ex hypothesi* is unconscious, is also accessible, just that the pathway to access is longer. That is, it might be thought that we can really only talk in terms of *degrees* of accessibility, where explicitly represented information is *maximally* accessible. However, the status of explicitness is special, since for potentially explicit information to have effects on other networks it must *first be rendered explicit*. *Qua* potentially explicit, such information has no effects outside the network in which it is encoded.

8

The Marks of the Mental

Let me summarise the preceding chapters. My project has been to examine the prospects for a connectionist theory of consciousness. I motivated the project by suggesting that connectionism has the potential to shed new light on a debate concerning the fundamental criteria of mentality. I'll shortly turn to this issue, but clearly the implications of connectionism for the study of consciousness are of independent interest, especially in view of the fact that connectionism is in a position to offer a vehicle theory of consciousness.

The two current candidates for a computational account of human cognition (classicism and connectionism) agree that human thought consists in semantically coherent processes defined over in-the-head representations (an idea they borrow from the folk), but differ concerning both the nature of these representations, and the kinds of processes in which they are implicated. I have characterised the disagreement using Dennett's (1982) taxonomy of styles of mental representation, and suggested that in PDP systems, unlike digital systems, the distinction between potentially explicit and tacit representation collapses. This amounts to the familiar claim that in PDP networks the structures that subserve the storage of information are the very structures that determine the course of processing. For this reason connectionists have some serious purchase on the global sensitivity of thought, a point that is now widely recognised. What is not so widely recognised, however, is that the unique characteristics of connectionist representation and processing also permit connectionists to advance a vehicle theory of consciousness: a theory which seeks to identify phenomenal experience with the explicit representation of information in the brain. Classicists are not in a position to offer this sort of theory, because the computational resources available in a digital system commit them to the dissociation thesis.

The specific proposal I considered in the last chapter is that the elements of phenomenal experience are identical to stable patterns of activation in neurally realised PDP networks. There is much to recommend this proposal. Our examination of phenomenal experience in Chapter 3 revealed that instantaneous consciousness is densely polyphonic – an amalgam of phenomenal elements, both sensory and non-sensory – and that ongoing experience is woven from a great many semi-independent strands. Research in the neurosciences suggests that this rich phenomenal fabric is generated at a multitude of consciousness-making sites scattered throughout the brain. A multi-track model of consciousness thus seems to be required. Such a model is consistent with the unity of consciousness, despite first appearances, because "unity" need not be conceived literally in terms of oneness – consciousness is not one thing; it is a bundle of co-occurrent, albeit tightly co-ordinated elements. It is precisely this picture of consciousness that emerges when we identify phenomenal experience with stable patterns of activation in neurally realised PDP networks. Consciousness is polyphonic because it consists of a great many stable activation patterns, each of which contributes distinct contents to the phenomenal field; the multiplicity of these patterns, considered as the products of physically distinct neural networks, renders the connectionist theory a multi-track model of consciousness; and the massive connectivity of the brain acts to bring the various parts of experience into coherence. We've also seen that the connectionist vehicle theory of consciousness makes sense of the sensory/non-sensory

hierarchy discernible in experience, and has the explanatory resources to account for the enormous diversity of experience.

The most serious obstacle facing the connectionist vehicle theory of consciousness is the current consensus in favour of the dissociation thesis: the claim that information can be explicitly represented in the brain without entering consciousness. I considered various kinds of direct evidence for this thesis in Chapter 5, and found most of it to be methodologically flawed in one respect or another. Seemingly significant effects tend to disappear when attempts are made to replicate them under more stringent conditions, or else prove to be the result of simple, unforseen conscious processes. A great many studies support the dissociation thesis *in*directly, by way of an inference to the best explanation, but this form of argument is always vulnerable to alternative explanations of the relevant phenomena. In Chapter 7 I indicated the flavour of connectionist explanations for a wide range of phenomena, including: automatisation; intuitive and symbolic reasoning; perceptual processing; and working memory. While often conjectural, these explanations show that connectionism has the potential to dispense with the dissociation thesis without losing its grip on the traditional targets of cognitive science.

For all of these reasons I believe the connectionist vehicle theory of consciousness is a worthy object of further investigation and elaboration. It challenges orthodoxy regarding both the structure of experience, and the nature of the (unconscious) processes that subserve cognition, but in so doing brings into clearer focus the issues with which any theory of consciousness much contend.

Finally, then, to the mark of the mental. In the introduction I distinguished two possible criteria for mentality: the consciousness criterion, which treats phenomenal properties as the ultimate mark of the mental; and the intentionality criterion, which puts computation first by linking mentality to the representation of information. The latter allows for unconscious mental states, i.e., mental representations whose contents don't figure among the elements of phenomenal experience. The former permits mental states that are non-representational (phenomenal states that are not "about" anything), but cannot countenance unconscious mentality. However, as Rosenthal points out, there is an important caveat of which the consciousness criterion can avail itself:

> Perhaps dispositional or cognitive states exist that are not conscious, but nonetheless count as mental states. But if so, such states would be derivatively mental, owing their mental status solely to their connection with conscious states. (1986, p.462)

The idea here is to introduce a non-occurrent sense in which consciousness is the mark of the mental. Certain brain states could, for example, count as mental because of their *disposition to generate* conscious states. This is perhaps what Searle is getting at when he claims that the "*ontology of the unconscious consists in objective features of the brain capable of causing subjective conscious thoughts*" (1989, p.202, italics in the original, see also his 1992, ch.7).

Given the enormous part played by unconscious representations in conventional accounts of cognition, those who are attracted to the computational theory of mind have generally felt compelled to adopt the intentionality criterion of mentality. But, in light of the approach to phenomenal experience I've canvassed in this thesis, such a response now appears premature. Theorists are *not* faced with a choice between adopting the intentionality criterion, or taking on the burden of articulating a non-computational account of cognition. Rather, they are faced with two varieties of computational theory: classicism, which does seem to be committed to the intentionality criterion; and connectionism, which has the potential to put quite a different spin on the relationship between consciousness and mental representation.

According to the connectionist vehicle theory of consciousness, phenomenal experience is identical to the explicit representation of information in the brain. If this account proves workable then consciousness is far more fundamental to cognition than most theorists have been willing to concede – the very states (activation pattern representations) that constitute the brain's occurrent response to prevailing conditions, and which support robust inter-network communication, are none other than the elements of conscious experience. While there is still a crucial role for unconscious processes in this account, such processes are not defined over explicit representations. Rather, explicit representations, in the form of stable patterns of activation in neurally realised PDP networks, are the *products* of such processes. We may still speak of unconscious information in this context – it is entirely natural to regard the brain as storing an enormous amount of information – but such information is *inexplicitly* represented. It doesn't take the form of discrete physical packages, because superpositional encoding is the order of the day for brain-based storage. Consequently, unconscious information has something of a derived status in the present account. Unconscious contents are individuated by reference to a dispositional property of neural networks: their capacity to generate a whole range of explicit representations in response to suitable input. Since explicit representations are the elements of consciousness, this reduces unconscious mental states to brain states that are apt for the production of conscious contents.

All of this should put one in mind of Rosenthal's take on the consciousness criterion of mentality. As he remarks, while the consciousness criterion treats phenomenal properties as the defining mark of the mental, it appears to allow for the existence of (unconscious) cognitive states that are derivatively mental, by virtue of their connection with conscious states. And this seems to be the status of unconscious representations within the connectionist vehicle theory of consciousness. Those network states responsible for unconscious processing are recognised by virtue of their *disposition to generate* explicit (and hence conscious) states.

But the dichotomy, either intentionality or consciousness as the mark of the mental, is too simple to do justice to the connectionist vehicle theory of consciousness. Certainly, the place of conscious experience in the scheme of things is radically altered by that account. However, it would be a mistake to see this as a victory for the consciousness criterion. What really follows, if one takes the connectionist vehicle theory of consciousness seriously, is a surprising convergence of intentionality and consciousness. It is true that unconscious contents are identified with the conscious states a network has the capacity to generate, but these states are stable activation patterns – the focal representational structures in the connectionist scheme. Consciousness and intentionality come together here in a physical kind that is at once representational and phenomenal in character: aboutness becomes the very stuff of conscious experience. This suggests that the search for a single criterion of mentality is a dubious enterprise. Intentional and phenomenal properties are equally potent marks of the mental.

References

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

Anderson, J.R. (1995) *Cognitive Psychology and its Implications*, 4th ed.

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

Baars, B.J. (1988) *A Cognitive Theory of Consciousness*, CUP

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

Bach-y-Rita, P (1972) Brain Mechanisms in Sensory Substitution, Academic Press

Baddeley, A.D. (1986) *Working Memory*, OUP

Bechtel, W. (1987) Connectionism and the philosophy of mind: An overview. *The Southern Journal of Philosophy* **26**:17-41 (Supplement)

Bechtel, W. (1988) *Philosophy of Mind: An Overview for Cognitive Science*, Lawrence Erlbaum

Birkhoff, G. & MacLane, S. (1977) *A Survey of Modern Algebra*, 4th ed., Macmillan Publishing Co.

Bisiach, E. (1992) Understanding consciousness: Clues from unilateral neglect and related disorders. In A. Milner and M. Rugg (eds.) *The Neuropsychology of Consciousness*, Academic Press

Block, N. (1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* **18**:227-47

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

Broadbent, D.E. (1958) *Perception and Communication*, Pergamon Press

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

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

Chase, W.G. & Simon, H.A. (1973) Perception in chess. *Cognitive Psychology* **4**:55-81

Cherry, E.C. (1953) Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of* America **25**:975-9

Chomsky, N. (1959) Review of B. F. Skinner's *Verbal Behavior*. *Language* **35**:16-58

Chomsky, N. (1980) Rules and Representations. *Behavioral and Brain Sciences* **3**:1-61

Churchland, P.M. (1979) *Scientific Realism and the Plasticity of Mind*, Cambridge University Press

Churchland, P.M. (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* **78**:67-90

Churchland, P.M. (1989) *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press

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

Churchland, P.S. & Sejnowski, T.J. (1989) Neural representation and neural computation. In L.Nadel, L.A.Cooper, P.Culicover & R.H.Harnish (eds.) *Neural Connections and Mental Computations*, MIT Press

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

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

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

Clark, A. (1993b) Minimal Rationalism. *Mind* **102**:587-610

Clark, A. (1997) *Being There: Putting Brain, Body, and World Together Again*, MIT Press

Clark, A. & Karmiloff-Smith, A. (1993) The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language* **8**(4):487-519

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

Copeland, J. (1993) *Artificial Intelligence: A Philosophical Introduction*, Blackwell

Corteen, R.S. (1986) Electrodermal responses to words in an irrelevant message: A partial reappraisal. *Behavioral and Brain Sciences* **9**:27-8

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

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

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

Cussins, A. (1990) The connectionist construction of concepts. In M.Boden (ed.) *The Philosophy of Artificial Intelligence*, OUP

Cytowic, R.E. (1993) *The Man Who Tasted Shapes*, Abacus

Davidson, D. (1974) Belief and the basis of meaning. *Synthese* **27**:309-23

Dehaene, S. & Changeux, J-P. (1989) A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience* **1**:244-61

De Houwer, J., Hendrickx, H., Baeyens, F. & Van Avermaet, E. (1993) *Hidden covariation detection might be very hidden indeed*, Manuscript submitted for publication.

Dellarosa, D. (1985) A History of Thinking. In R.J.Sternberg & E.E.Smith (eds.) *The Psychology of Human Thought*, CUP

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

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

Dennett, D.C. (1987) *The Intentional Stance*, MIT Press

Dennett, D.C. (1991a) Real patterns. *Journal of Philosophy* **88**:27-51

Dennett, D.C. (1991b) *Consciousness Explained*, Little, Brown and Company

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

Dennett, D.C. (1996) Consciousness: More like fame than television (in German trans.). In C.Maar, E.Pöppel & T.Christaller (eds.) *Die Technik auf dem Weg zur Seele*, Rowuhlt

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

Devitt, M. & Sterelny, K. (1987) *Language and Reality: An Introduction to the Philosophy of Language*, Blackwell

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**:875-87

Dretske, F. (1988) *Explaining Behavior: Reasons in a World of Causes*, MIT Press

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

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

Dreyfus, H.L. & Dreyfus, S.E. (1986) How to stop worrying about the frame problem even though it's computationally insoluble. In Z.Pylyshyn (ed.) *The Robot's Dilemma*

Dulany, D.E. (1991) Conscious Representation and Thought Systems. In R.S.Wyer Jr. & T.K.Srull (eds.) *Advances in Social Cognition IV*, Erlbaum

Dulany, D.E. (1996) Consciousness in the Explicit (Deliberative) and Implicit (Evocative). In J.Cohen & J.Schooler (eds.) *Scientific Approaches to Consciousness* (in press), Erlbaum

Dulany, D.E. & Poldrack, R.A. (1991) Learned covariation: Conscious or unconscious representation? Paper read at the Psychonomic Society, San Francisco, CA.

Elman, J.L. (1989) Representation and Structure in Connectionist Models. CRL Technical Report

Elman, J.L. (1990) Finding Structure in Time. *Cognitive Science* **14**:179-211

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

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

Fodor, J.A. (1975) *The Language of Thought*, Thomas Crowell

Fodor, J.A. (1978) Three cheers for propositional attitudes. In E.Cooper & E.Walker (eds.) *Sentence Processing*, Lawrence Erlbaum Associates

Fodor, J.A. (1981) *RePresentations*, MIT Press

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

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

Fodor, J.A. (1991) Replies. In B. Loewer & G. Rey (eds.) *Meaning in Mind: Fodor and His Critics*, Blackwell

Fodor, J.A. (1997) Connectionism and Systematicity. *Cognition* **62**:109-19

Fodor, J.A. & McLaughlin, B.P. (1990) Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* **35**:183-204

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

Gardner, H. (1985) *The Mind's New Science: A History of Cognitive Revolution*, Basic Books

Geschwind, N. (1990) Specializations of the Human Brain. Reprinted in R.R.Llinás (ed.) *The Workings of the Brain: Development, Memory, and Perception*, W.H.Freeman and Company

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

Goldman-Rakic, P.S. (1990) Parallel systems in the cerebral cortex: The topography of cognition. In M.Arbib & J.Robinson (eds.) *Natural and Artificial Parallel Computation*, MIT Press

Goldman-Rakic, P.S. (1992) Working memory and the mind. *Scientific American* **267**:72-9

Hadamard, J. (1949) *The Psychology of Invention in the Mathematical Field*, Princeton University Press

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

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

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

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

Haugeland, J. (1987) An overview of the frame problem. In Z.W.Pylyshyn (ed.) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Ablex

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

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

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

Horgan, T. & Tienson, J. (1996) *Connectionism and the Philosophy of Psychology*, MIT Press

Horgan, T. & Woodward, J. (1985) Folk psychology is here to stay. *The Philosophical Review* **94**:197-226

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

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

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

James, W. (1892) *Psychology, Briefer Course*, Holt, ch.11, pp.152-175, reprinted in A.J.Reck (1967) *Introduction to William James*, Indiana University Press

Johnson-Laird, P.N. (1988) A taxonomy of thinking. In R.J.Sternberg & E.E.Smith (eds.) *The Psychology of Human Thought*, Cambridge University Press

Johnson-Laird, P.N. (1993) *The Computer and the Mind*, 2nd ed., Fontana Press

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

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

Jordan, M.I. (1989) Serial order: A parallel distributed processing approach. In J.L.Elman & D.E.Rumelhart (eds.) *Advances in Connectionist Theory*, Erlbaum

Kinsborne, M. (1988) Integrated field theory of consciousness. In A.Marcel & E.Bisiach (eds.), *Consciousness in Contemporary Science*, Clarendon Press

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

Kolers, P.A. (1972) *Aspects of Motion Perception*, Pergamon Press

Kosslyn, S.M. (1980) *Image and Mind*, Harvard University Press

Kosslyn, S.M. (1981) The medium and the message in mental imagery. In N.Block (ed.) *Imagery*, MIT Press

Kulpe, O. (1964) The modern psychology of thinking. In J.Mandler & G.Mandler (eds.) *Thinking: From Association to Gestalt*, Wiley

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

Land, E. (1977) The retinex theory of color vision. *Scientific American* December: 108-28

Lashley, K. (1956) Cerebral organisation and behavior. In H.Solomon, S.Cobb & W.Penfield (eds.) *The Brain and Human Behavior*, Williams & Wilkins Press

Lecours, A.R. & Joanette, Y. (1980) Linguistic and Other Aspects of Paroxysmal Aphasia. *Brain and Language* **10**:1-23

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**:251-78

Lewicki, P. (1986) Processing information about covariation that cannot be articulated. *Journal of Experimental Psychology: Learning , Memory, and Cognition* **12**:135-46

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

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

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

Lockwood, M. (1989) *Mind, Brain and the Quantum: The Compound 'I'*, Blackwell

Logan, G.D. (1988) Toward an instance theory of automatization. *Psychological Review* **95**(4):492-527

Logan, G.D. (1990) Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology* **22**:1-35

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

Mandler, G. (1975) Consciousness: Respectable, useful, and probably necessary. In R.Solso (ed.) *Information Processing and Cognition*, Erlbaum

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

Mandler, G. & Shebo, B.J. (1982) Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General* **111**:11-22

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

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

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

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

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

Marr, D. & Poggio, T. (1976) Co-operative computation of stereo disparity. *Science* **194**:283-287

Mates, B. (1965) *Elementary Logic*, OUP

Mayer, R.E. (1983) *Thinking, Problem Solving, Cognition*, W.H.Freeman and Company

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

McClelland, J.L., Rumelhart, D.E. & Hinton, G.E. (1986) The appeal of parallel distributed processing. In Rumelhart & McClelland 1986

Metzinger, T. (1995) Faster than thought: Holism, homogeneity and temporal coding. In T.Metzinger (ed.) *Conscious Experience*, Academic Imprint

Miller, G.A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review* **63** (2):81-97

Miller, G.A. & Buckhout, R. (1973) *Psychology: The Science of Mental Life* (2nd ed.), Harper and Row

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

Minsky, M. (1985) *The Society of Mind*, Picador

Moray, N. (1959) Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology* **11**:56-60

Moray, N. & O'Brien, T. (1967) Signal-detection theory applied to selective listening. *Journal of the Acoustical Society of America* **42**:765-72

Mowbray, G.W. (1964) Perception and retention of verbal information presented during auditory shadowing. *Journal of the Acoustical Society of America* **36**:1459-64

Neisser, U. (1967) *Cognitive Psychology*, Appleton-Century-Crofts

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

Newell, A. (1980) Physical Symbol Systems. *Cognitive Science* **4**:135-83

Newell, A. & Simon, H.A. (1972) *Human Problem Solving*, Prentice-Hall

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

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

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

O'Brien, G. (1993) *The Computational Theory of Mind: An Exploration of the Conceptual Foundations of Cognitive Science*, Doctoral Thesis, New College, Oxford

Parfit, D. (1984) *Reasons and Persons*, OUP

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

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

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

Perenin, M.T. & Jeannerod, M. (1978) Visual function within the hemianopic field following early cerebral hemidecortication in man. I. Spatial localisation. *Neuropsychologia* **16**:1-13

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

Phaf, R.H. & Wolters, G. (1997) A constructivist and connectionist view on conscious and nonconscious processes. *Philosophical Psychology* **10**:287-307

Pinker, S. (1994) *The Language Instinct*, Penguin Books

Place, U.T. (1956) Is consciousness a brain process? *British Journal of Psychology* **47**:44-50

Poincaré, H. (1982) *The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method*, University Press of America

Pollack, I. (1952) The information of elementary auditory displays. *The Journal of the Acoustical Society of America* **24**(6):745-9

Posner, M.I. (1978) *Chronometric Explorations of Mind*, Lawrence Erlbaum Associates

Posner, M.I. & Boies, S.J. (1971) Components of attention. *Psychological Review* **78**:391-408

Posner, M.I. & Klein, R.M. (1973) On the functions of consciousness. In S.Kornblum (ed.) *Attention and Performance IV*, Academic Press

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

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

Pylyshyn, Z.W. (1984) *Computation and cognition: Toward a foundation for cognitive science*, MIT Press

Ramachandran, V.S. (1988) Perceiving shape-from-shading. *Scientific American* **259**:76-83

Ramachandran, V.S. & Anstis, S.M. (1986) The perception of apparent motion. *Scientific American* **254**:80-7

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

Rey, G. (1993) Sensational sentences. In M.Davies & G.W.Humphreys (eds.) *Consciousness: Psychological and Philosophical Essays*, Blackwell

Rich, E. & Knight, K (1991) *Artificial Intelligence*, 2nd ed., McGraw-Hill

Rock, I. (1983) *The Logic of Perception*, MIT Press

Rose, S. (1993) *The Making of Memory*, Bantam Books

Rosenthal, D.M. (1986) Two Concepts of Consciousness. *Philosophical Studies* **94**:329-59, reprinted in D. M. Rosenthal (ed.) (1991) *The Nature of Mind*, OUP

Rumelhart, D.E., Hinton, G.E. & McClelland, J.L. (1986) A general framework for parallel distributed processing. In Rumelhart & McClelland 1986

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning internal representations by error propagation. In Rumelhart & McClelland 1986

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

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

Ryle, G. (1949) *The Concept of Mind*, Hutchinson

Sacks, O. (1985) *The Man Who Mistook his Wife for a Hat*, Picador/Pan Books

Sacks, O. (1995) *An Anthropologist on Mars*, Picador/Pan Books

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

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

Searle, J.R. (1980) Minds, brains and programs. *Behavioral and Brain Sciences* **3**:417-57

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

Searle, J.R. (1989) Consciousness, Unconsciousness and Intentionality. *Philosophical Topics* **17**:193-209

Searle, J.R. (1990) Is the brain's mind a computer program? *Scientific American* **262**:20-5

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

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

Seyfarth, R.M. & Cheney, D.L. (1984) The natural vocalizations of non-human primates. *Trends in Neuroscience* **7**:66-73

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

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

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

Smart, J.J.C. (1959) Sensations and brain processes. *Philosophical Review* **68**:141-56

Smith, B.C. (1982) Semantic attribution and the formality condition. Paper presented at the Eighth Annual Meeting of the Society for Philosophy and Psychology, University of Western Ontario

Smith, S.M. (1995a) Getting into and out of mental ruts: A theory of fixation, incubation and insight. In R.J.Sternberg & J.E.Davidson (eds.) *The Nature of Insight*, MIT Press

Smith, S.M. (1995b) Fixation, incubation, and insight in memory and creative thinking. In S.M.Smith, T.B.Ward & R.A.Finke (eds.) *The Creative Cognition Approach*, MIT Press

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

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

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* **11**:1-74

Smythies, J.R. (1994) Requiem for the identity theory. *Inquiry* **37**:311-29

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

Stich, S. (1978) Beliefs and subdoxastic states. *Philosophy of Science* **45**:499-518

Stillings, N.A., Weisler, S.E., Chase, C.H., Feinstein, M.H., Garfield, J.L. & Rissland, E.L. (1995) *Cognitive Science: An Introduction*, MIT Press

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

Tortora, G. & Anagnostakos, N. (1987) *Principles of Anatomy and Physiology* (5th ed.), Harper and Row

Turing, A. (1937) On computable numbers, with an application to the *entscheidungsproblem*. *Proceedings of the London Mathematical Society* (Series 2) **42**:230-65

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

Tye, M. (1995) Blindsight, orgasm, and representational overlap. *Behavioral and Brain Sciences* **18**:268-9

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

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

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

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

van Gelder, T. (1991) What is the 'D' in 'PDP'? A survey of the concept of distribution. In W.Ramsey et al (eds.) *Philosophy and Connectionist Theory*, Erlbaum

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

Vokey, J.R. & Brooks, L.R. (1992) Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**:328-44

Vygotsky, L.S. (1962) *Thought and Language*, MIT Press

Wallas, G. (1931) *The Art of Thought*, Cape

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

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

Weiskrantz, L., Warrington, E.K., Sanders, M.D. & Marshall, J. (1974) Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain* **97**:109-28

Wilson, M.E. (1968) The detection of light scattered from stimuli in impaired regions of the visual field. *Journal of Neurology, Neurosurgery and Psychiatry* **31**:509-13

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

Zipser, D. (1991) Recurrent network model of the neural mechanism o short-term active memory. *Neural Computation* **3**:178-9