Consciousness Distributed

Robert West

Contents

1. Wittgenstein’s Challenge

> How does the philosophical problem about mental processes and states and about behaviourism arise? - The first step is the one that altogether escapes notice. We talk of processes and states and leave their nature undecided. Sometimes perhaps we shall know more about them - we think. But that is just what commits us to a particular way of looking at the matter. For we have a definite concept of what it means to learn to know a process better. (The decisive movement in the conjuring trick has been made, and it was the very one that we thought quite innocent.) - And now the analogy which was to make us understand our thoughts falls to pieces. So we have to deny the yet uncomprehended process in the yet unexplored medium. And now it looks as if we had denied mental processes. And naturally we don’t want to deny them.[1]

Here Wittgenstein has articulated the central problem in the philosophy of mind. Philosophers approach the mind with a false picture of its nature. It is this: the mental is seen as consisting of discrete, isolable entities. Mental states are constituted by a certain arrangement of these entities, and mental processes are successions of such states. This leads us to talk of beliefs, desires, pains, and so on, as though they were distinct mental objects, directly involved in the causation and explanation of behaviour.

This picture of the mental in terms of discrete entities, states, and processes is a presupposition of the realist positions that have been proposed in the course of modern philosophy of mind from Cartesian dualism, through to the various forms of functionalism. Furthermore, the legitimacy of realism about propositional attitudes rests on this picture, at least within the framework of the language of thought hypothesis.[2] According to this hypothesis mental content is characterised in sentential terms (using the philosophical paradigm of propositional attitudes) and is held to be represented in the brain by a syntactic system. Thus in this case the discrete mental entities are the symbols of a syntactic system, which determines their lawful combinations and causal roles within the cognitive environment. Within this paradigm understanding a domain is considered to be a matter of possessing a theory of that domain, where this theoretical knowledge is sentential in form.

I shall argue that, as Wittgenstein asserted, to think of the mental in terms of discrete entities, states, and processes is to misconceive the fundamental nature of the mental.[3] Much of Wittgenstein’s later philosophy can be interpreted as an attack on this false picture. He realised that traditional philosophy has generated an over-intellectualised account of human cognition, which is catalysed by misunderstandings about language:

> When we are worried about the nature of thinking, the puzzlement which we wrongly interpret to be one about the nature of a medium is a puzzlement caused by the mystifying use of our language.[4]

When we consider the nature of propositional attitudes, or sensation, our forms of language tempt us to think that we are dealing with entities on the same model as physical entities. The lesson to be taken from Wittgenstein’s philosophy is that such an approach leads to a distinction between purely exterior behaviour and purely interior mental phenomena, which immediately become problematic. We are forced to label them as intrinsic, or by some similar token as private. This puts them beyond the reach of third person observation, leading to problems about the existence of other minds and of the external world.

These misconceptions arise because of the enormous and unparalleled role language plays in *human* conscious life. Yet in evolutionary terms it is a comparatively modern innovation, and as such is more likely to be an optional extra, rather than part of the basic configuration of consciousness. A

[1] Wittgenstein, L. 1953. §308.

[2] See Fodor, J. 1975.

[3] My argument involves two elements: (1) the mental cannot be reduced to either physical types, or isolable physical tokens, and (2) the type/token notion derived from physical objects cannot be applied to the mental, which has a different ontological status.

[4] Wittgenstein, L. 1958. p.6.

plethora of animals share mental features with us, without the need for language. This indicates that language should play only a marginal role in the explanation of the *fundamental* nature of mental life and consciousness, particularly of perception and content.

This suggests that a pre-linguistic model is needed. However, for Wittgenstein philosophy was a purely descriptive enterprise, whose task is to dissipate these linguistic confusions, and not to provide explanations of internal cognitive phenomena. Since I do not accept this limitation, I wish to take Wittgenstein's comments as a challenge, provoking a fresh examination of the conceptual landscape of the mental. By redefining the mental in terms of a pre-linguistic, distributed model, inspired by connectionism,[5] I shall attempt to demonstrate that Wittgenstein's criticisms of the discrete entity model are correct, but that we can deal positively with the mental while avoiding his arguments against theorising about internal processes.

Despite its problems, the discrete entity picture has remained the dominant framework for the philosophy of mind and cognitive science, simply because there is no fully developed alternative model to compete with it. I believe that connectionism provides the raw materials for just such an alternative model. The truth of connectionism is a matter for much future empirical research, but in this thesis I will assume that it is at least an approximation to the truth, in order to explore some of its possible philosophical consequences, and to produce a speculative sketch of a distributed theory of mind.

Section 2 begins this task with an explanation of Dennett's multiple drafts model of consciousness. This provides a useful means of breaking out of old conceptual habits, because it attacks one of the most seductive manifestations of the discrete entity view: Cartesian materialism. This is what you get when you jettison the metaphysical baggage of Cartesian dualism, whilst still clinging to the notion of a "finishing post" in the brain which determines the exact moment at which something becomes conscious. Dennett's intuition pumps and arguments against this position, and his alternative model, are important in two ways. Firstly they dissolve the central point, illustrating how consciousness can be considered a distributed, multi-track phenomenon. Secondly they demonstrate the counter-intuitive vagueness and indeterminacy of consciousness; in some cases there just is no fact of the matter about whether something reaches consciousness or not, making it more amenable to a distributed conceptualisation. Whilst the multiple drafts model avoids many of the errors brought about by the discrete entity picture - due to Dennett's instrumentalism - I shall argue that with regard to phenomenal consciousness it is too verificationist and eliminativist in flavour, and thus needs substantial amendment.

Section 3 introduces the main features of connectionism, and details their advantages over conventional symbolic approaches in producing this more plausible account of the mind. The essential difference is captured by the notion of semantic transparency, where a system has this property "if there is a neat mapping between states that are computationally transformed and semantically interpretable bits of sentences."[6] Conventional symbolic approaches are semantically transparent, whilst connectionism is semantically opaque. I shall argue that cognition is not semantically transparent, and thus that the language of thought hypothesis is false. The contrasting connectionist approach involves a characterisation of knowledge in terms of the emergent skills and abilities exhibited by networks when considered as complete systems, rather than in terms of sentences. Thus connectionism offers a way of accommodating Ryle's distinction between knowing how and knowing that, and Wittgenstein's anti-linguistic account of understanding, whilst also being evolutionarily plausible.

Section 4 answers an important question raised by these conclusions, about the status and authority verbal reports of mental states really have if they are not about particular mental entities. I suggest how connectionist ideas can provide an account of folk psychology which is sympathetic to

[5] This term is used in different ways by different people, but roughly it describes an approach to cognitive modelling and artificial intelligence that utilises many simple units which are richly interconnected. See section 3 for a more complete account.

[6] Clark, a. 1989. p.2.

our intuitions. I propose a new kind of relationship between the propositional attitude/linguistic level, and the level of neurophysiological processes. Intentional phenomena are unmysteriously emergent properties, the product of a mass of distributed neural processing, whose component parts are not susceptible to semantic evaluation. This creates a principled basis for non-reductive supervenience of the mental.

Once this stage-setting has been done it will then be possible, in section 5, to provide a connectionist account of phenomenal consciousness. This model has its roots in Dennett's multiple drafts theory, but does not "feign anaesthesia" in the way that I shall argue Dennett has done, despite his protestations to the contrary. By restructuring the debate according to connectionist principles, describing mental phenomena in terms of distributed representations the inner/outer, behavioural disposition/qualia distinctions lose their utility, as Wittgenstein has demanded, and many of the paradoxes surrounding phenomenal consciousness can be resolved.

2. The Ghost in the Joycean Machine

The idea of a finishing point in the head exercises a surreptitious influence on the philosophical view of the mental, and is one of the principal ways in which the discrete entity picture is articulated in the philosophy of mind. Within the bounds of normal human activity it makes perfect sense to conceive of a person as the point where all incoming information is received, and from which all behaviour flows. On this view there is a definitive finishing post for sensory information, which determines when something is present in consciousness; this is the place that Dennett calls the Cartesian Theatre:

> The Cartesian Theatre is a metaphorical picture of how conscious experience must sit in the brain. It seems at first to be an innocent extrapolation of the familiar and undeniable fact that *for everyday, macroscopic time intervals*, we can indeed order events into the two categories "not yet observed" and "already observed." We do this by locating the observer at a point and plotting the motions of the vehicles of information relative to that point. But when we try to extend this method to explain phenomena involving very short time intervals, we encounter a logical difficulty: If the "point" of view of the observer must be smeared over a rather large volume in the observer's brain, the observer's own subjective sense of sequence and simultaneity must be determined by something other than "order of arrival," since order of arrival is incompletely defined until the relevant destination is specified.[7]

Thus it is phenomena on the microscopic time scale that challenge the notion of the Cartesian theatre. Dennett brings many experimental findings to bear on this point, but it is best illustrated by his 'woman in glasses' example. Suppose I now seem to have a memory that a woman with long hair and glasses has just dashed by. In actuality, however, the woman who went past had no glasses, but I have confused her in my mind with a woman I saw earlier in the week who had short hair and glasses. Within the Cartesian framework there are two possible ways to describe what is going on here, either an Orwellian revision, or a Stalinesque fabrication, in Dennettian parlance. In the Orwellian version I really did see the woman without glasses, but a split second after this was conscious the memory was erased and replaced by the false memory of a woman with long hair and glasses. For the Stalinesque version all that really was presented for conscious was the long haired woman in glasses, the glasses being edited in before the image reached consciousness. According to Dennett these two hypotheses cannot be distinguished either 'from the outside', or 'from the inside.' In both cases the subject's reports will be the same, and any neurophysiological evidence will beg the question by assuming a finish line:

> Here the distinction between perceptual revisions and memory revisions that works crisply at other scales is no longer guaranteed to make sense. We have moved into the foggy area in which the subject's point of view is spatially and temporally smeared, and the question *Orwellian or Stalinesque*? loses its force.[8]

In order to account for this situation Dennett proposes his multiple drafts model of consciousness. The name comes from an analogy with modern publishing, where there may be many differing versions of a work in circulation prior to its official publication, due to the technology now available. As many of the people that matter have already read a copy, the official version becomes increasingly unimportant. As it is with publishing, so it is with the mind, according to Dennett:

> All variety of thought or mental activity are accomplished in the brain by parallel, multi-track processes of interpretation and elaboration of sensory inputs. Information entering the nervous system is under continuous 'editorial revision.'[9]

According to this model processing occurs in many streams throughout the brain, with each stream undergoing many "additions, incorporations, emendations and overwritings of content." Once a

[7] Dennett, D. C. 1991. p.107.
[8] Ibid. p.119.
[9] Ibid. p.111.

discrimination has been made in a particular portion of the brain that is all there is, the information plays its part in the ongoing processing without having to be replayed for some sort of master system:

> These spatially and temporally distributed content-fixations in the brain are precisely locatable in both space and time, but their onsets do not mark the onset of consciousness of their content. It is always an open question whether any particular content thus discriminated will eventually appear as an element in conscious experience, and it is a confusion to ask *when it becomes conscious*.[10]

What we are conscious of, according to Dennett, cannot be determined independently of the probes used to evoke a narrative (either a linguistic internal representation, or a verbal report). A quick probe during a given task may produce an "incomplete" draft, and effect the "flow" of the multiple streams. A late probe might produce no narrative at all, as the information is no longer active within the system. It seems curious to define consciousness in terms of narrative precipitation, yet this is a consequence of Dennett's approach, which he terms heterophenomenology. This involves treating the subject's verbal report as a (theorist's) fiction, by which it is meant that:

> [The] method neither challenges nor accepts as entirely true the assertions of subjects, but rather maintains a constructive and sympathetic neutrality, in the hopes of compiling a definitive description of the world according to the subjects.[11]

The idea is to account for the subject's reports in a way that does not make any theoretical presuppositions. Dennett wants to explain why there should *seem* to be phenomenology without there actually being any. All that occur in the brain are content-fixations, distributed throughout the system, whose only effect is to inform other processes with their content. Some of this activity will lead to linguistic utterances, either public or internal. It is this text that creates the "benign illusion" that there is actual phenomenology, instead of heterophenomenology. Dennett robustly defends this verificationist approach:

> Some thinkers have their faces set so hard against "verificationism" . . . that they want to deny it even in the one arena where it makes manifest good sense: the realm of subjectivity. . . . [They object that] "Just because you can't tell, by your preferred ways, whether or not you were conscious of x, that doesn't mean you weren't. Maybe you were conscious of x but just can't find any evidence for it!" Does anyone, on reflection, really want to say that? Putative facts about consciousness that swim out of reach of both "outside" and "inside" observers are strange facts indeed.[12]

The apparent plausibility of Dennett's argument rests on an equivocation between *access* and *phenomenal* consciousness. Phenomenal consciousness is difficult to define precisely, but it is intuitively captured by Nagel's phrase: *what it is like* to be a subject of experience. Block describes a state as access conscious if:

> in virtue of one's having the state, a representation of its content is (a) inferentially promiscuous, i.e. freely available as a premise in reasoning, and (b) poised for rational control of action, and (c) poised for rational control of speech.[13]

Although this definition mentions mental states, and thus involves itself in the picture I am arguing against, it still gives a feel for the sort of distinction that needs to be drawn. In many ways the human mind *can* be considered as an information processing device (and thus as access conscious) and here Dennett's ideas are very useful in accounting for experimental findings from psychology. Yet this is just a way of interpreting the brain, an approach which renders the mind more amenable to scientific study. It ignores important features of the mental which a philosophical account of consciousness cannot afford to miss without becoming spurious. In some cases there just is no

[10] Ibid. p.113.
[11] Ibid. p.83.
[12] Ibid. p.132.
[13] Block, N. 1994. p.214.

matter of fact about phenomenal consciousness, and in these situations Dennett's access conscious (multiple drafts) model can be brought to bear as an explanation of subject's reports. However, the multiple drafts model *is* incomplete because it only captures the *structure* of some elements of human behaviour and cognitive processing. It provides a good model of subject's reports for phenomena, particularly those experienced over a small time frame, and of the information processing that underlies this. Thus it can explain how definite verbal reports arise out of the apparently multifarious forms of activity that occur in the brain, yet it ignores the nature of these internal phenomena themselves. Dennett uses the computer as a metaphor, arguing that there is a Joycean machine, a serial, virtual programme running in the massively parallel brain which gives rise (through the operation of the multiple drafts model) to:

> something rather like a narrative stream or sequence, which can be thought of as subject to continual editing by many processes distributed around in the brain, and continuing indefinitely into the future.[14]

This Joycean machine, however, is just as ghostly as that which Ryle attacked if it is not properly specified. The ambiguity allows Dennett to make his position seem less behaviouristic and eliminativistic than it actually is. Thus Dennett's account fails because it places too much stress on the production of a *text*; but there are many aspects of phenomenal consciousness that cannot be captured by a text (where this is understood as a symbolic representation). This objection does not imply some sort of mysterious, ineffable qualia, rather it respects the minor role that language plays in underlying cognitive processing. Dennett admits to the role of internal processes in producing subjects reports about phenomenology, but by underspecifying the nature of these internal processes, his is a model of access consciousness only.

The phenomenal world of experience is by no means as determinate, immediate and incorrigible as many philosophers have assumed. Dennett is right to illustrate this aspect of consciousness, but he fails to acknowledge that while some cases may not admit of any positive fact of the matter concerning phenomenal consciousness, there are cases which are unproblematically definite. Thus instead of an absolute denial of phenomenal experience, we might talk in terms of a spectrum of instances, from brain events that are definitely unconscious, through the ambiguous cases, to the definite cases of phenomenal consciousness. I will call this the blurred boundaries model.

In order to see what is wrong with the multiple drafts model (and traditional philosophical views), it is necessary to look at Dennett's arguments for how it could seem that there is rich phenomenology where this "seems" is illusory:

> Suppose you walk into a room and notice that the wallpaper is a regular array of hundreds of identical - let's pay homage to Andy Warhol - photographic portraits of Marilyn Monroe. In order to identify a picture as a portrait of Marilyn Monroe, you have to foveate the picture: the image has to fall on the high-resolution fovea of your eyes . . . your *parafoveal* vision (served by the rest of the retina) does not have very good resolution. Yet we know that if you were to enter a room whose walls were papered with identical photos of Marilyn Monroe, you would "instantly" see that this was the case. You would see in a fraction of a second that there were "lots and lots of identical, detailed, focused portraits of Marilyn Monroe." Since your eyes saccade four or five times a second at most, you could foveate only one or two Marilyns in the time it takes you to jump to the conclusion *and thereupon to see* hundreds of identical Marilyns. We know that parafoveal vision could not distinguish Marilyn from various Marilyn-shaped blobs, but nevertheless, what you see is not wallpaper of Marilyn-in-the-middle surrounded by various indistinct Marilyn-shaped blobs.[15]

This is a good analysis of the way the human visual system works. It is more economical to leave the detail out their in the environment, where it is readily available when required, rather than producing an exhaustive "internal" copy. Thus in some instances the brain just represents *that* there is

[14] Dennett, D. C. 1991. p.113.

[15] Ibid. p.354.

something there, instead of representing the thing itself. In such cases our visual world gets to *seem* highly detailed because whichever area of the “visual field” we interrogate, by foveating on it, the detail is provided. The term visual field is in scare quotes because it implies an extremely pernicious metaphor for vision, namely that it is like being presented with pictures. It is presumed that when one does focus on a part of the visual field, it is just like concentrating on a particular patch of a picture. Yet this is a nonsensical position, for how can this picture have any existence when there is no means of rendering it in the brain? From what is a part of the picture constituted when I am not looking in that direction? The brain does not have the neural machinery capable of producing such a picture, because the parafoveal retina has poor resolution. Instead the brain just makes assumptions about what is there without “filling in” the parts of the picture that are not being viewed directly.

This may appear to be an implausible explanation at first, but it fits in with the psychological and physiological evidence about the functioning of the visual system. Vision does not just involve the seeing of areas of colour, of contours and textures; we see *objects*, where these are discrete entities, with their distinct form and function; seeing is seeing *as*. Thus there is an element of top down cognitive influence on what we see, where expectations and preconceptions influence the nature of the visual experience. Categorisation is an integral part of most visual experience. This is simply illustrated by illusions such as the duck-rabbit or the Necker cube (where a figure can switch between two differing visual interpretations), but it is normally presumed to work at a preconscious level. Talk of preconscious processing to some extent involves making presuppositions about the nature of the mental, which I would want to question, yet the basic phenomena remain to be explained regardless of the influence of theoretical language; vision is a much more fluid and cognitive process than many philosophers have assumed.

All of this suggests that the picture metaphor is inadequate and misleading, a consequence of the Cartesian Theatre where this picture would be presented. This lack of rigidity and specificity of vision begins to suggest ways in which the visual aspect of phenomenal consciousness might be less clear cut an issue than many philosophers have assumed. The idea of pictures in the head still exerts a surreptitious grip on the thinking of many, making the notion that visual experience is somehow ineffable, intrinsic and incorrigible seem necessary. This opens the way for a defence of qualia as something very special, yet when we begin to examine things more closely they do not appear so certain. Features that we thought were there in consciousness turn out to be absent.

This provides an opening for Dennett’s attack on the notion of qualia. The case of the Marilyns leads him to suggest that we are wrong about our visual experience, but the conclusion that he makes from this thought experiment is too strong. He wants to reduce all visual experience to representation *that* there is something there, rather than *actual* representation, but in his reliance on the creation of a text he fails to acknowledge important features of phenomenal consciousness. While it is true that in some cases it is not possible to say what the conscious experience consisted of (“Was that light red or blue? I get the feeling that it was red, but I’m not sure”) there are others where one definitely can say (“The Marilyn I am looking at *right now* is yellow and pink”). Hence the need for something like the blurred boundaries model to account for the apparent indeterminacy in visual experience, which does not assert that all seemings are false seemings. To have demonstrated the impossibility of the Cartesian Theatre is not to have proved the impossibility of phenomenal consciousness. There is another option, which is inspired by connectionism. I shall return to this problem in section 5 after discussing the significance of the connectionist model, and its impact upon the status of propositional attitudes.

3. Connectionism and Cognition

On the traditional view a mental state is a discrete entity, which is either conscious, or unconscious. In the way that consciousness is conceived there is no logical room for boarder line cases. Dennett's arguments indicate that consciousness (whether phenomenal or access) is more complex and sophisticated than this traditional picture allows. The orthodox conception of the mental fails to carve nature at its joints; it does not capture the way that the brain processes and represents sensory information.

Dennett goes some way to addressing this conceptual failure with his multiple drafts model, but I would argue that his model still places too much emphasis on the role of language in consciousness. He has not truly taken Wittgenstein's lesson to heart. A pre-linguistic theory of mental processing is needed if this deficiency is to be overcome. Connectionism (or parallel distributed processing, PDP) is just such a theory, and it can be contrasted with the linguistic approach in the form of conventional symbolic processing. The distinction between these two computational models can be made in terms of what Clark (1989) calls semantic transparency. A system is considered to be semantically transparent:

> if and only if it involves computational operations on syntactically specified internal states that (1) can be interpreted as standing for the concepts and relations spoken of in natural language (such items as "ball," "cat," "loves," "equals," and so on) and (2) these internal tokens recur whenever the system is in a state properly described by content ascriptions employing those words: the token is *projectible* to future cases. In short, a system is semantically transparent if there is a neat mapping between states that are computationally transformed and semantically interpretable bits of sentences.[16]

Thus, put simply, conventional (serial) computational models are semantically transparent, whilst connectionist models are semantically opaque. What I have called conventional computation is dependent on the concept of the Turing machine. This is a device which is capable of computing any computable function in a serial manner by simple symbol manipulation, and it is the inspiration for the digital computer. Conventional artificial intelligence (AI) argues that the mind works like such a device, and thus we get the idea of the mind as a semantically transparent system whose symbols constitute a language of thought. Such traditional models fall foul of Wittgenstein's arguments, because they take our notions of states and processes and attempt to reify them in the head.

Connectionism espouses an approach that is semantically opaque, so meanings are distributed within a parallel network of highly interconnected, but individually simple units. Such models are neurally inspired, but their exact structure and mathematical functions are unimportant. In what follows I will describe the behaviours of a range of different net architectures, but for the sake of simplicity I will ignore the technical details (no contemporary model will turn out to exactly mirror the way the brain achieves cognition, the importance of these networks is in the various emergent properties which they exhibit).

Most contemporary networks involve layers of units, which are analogous to neurones. Each unit in a given layer receives connections from all the units in the layer below, and sends connections to all the units in the layer above it. The connections are one-way, and each has a "weight" which determines the importance of that connection for the receiving unit (positive weights are excitatory, and negative weights are inhibitory). These units operate by summing the activity of their incoming connections using a fixed mathematical function. This then determines whether the unit "fires", usually by achieving a threshold level of activation. If the unit does fire it activates its connections to the units in the next layer. Networks usually have an input layer, which receives a pattern of stimulation representing the input from the outside world, patterns of sound, for example. There are then usually several hidden layers, so-called because they do not receive external input or produce external output, and then there is an output layer which displays the results of the networks

[16] Clark, A. 1989. p.2.

processing.

The behaviour of a network is determined by the weights of its connections, networks are not programmed, they are first assigned random weights and then given a period of learning. This involves the network producing answers for a set of inputs, which are then compared to the "correct" answer. If an incorrect answer is produced the weights of the connections are altered slightly to bring them closer to the correct answer, and over a large number of trials the network adopts the correct input/output mappings. Many different patterns of weights can produce networks with the correct input/output characteristics.

Such networks display an interesting range of abilities, which suggest that they may provide a better way to model human cognition than conventional AI. To account for knowledge of stereotypical items and situations conventional AI uses propositional schemata. A schema attempts to capture the relations within a given conceptual domain using propositions organised within an overall structure, with free argument places for specific pieces of information. These argument places have default settings representing the subject's assumptions about the conceptual domain. Thus Schank and Abelson (1977) posit a restaurant schema, which would be envisaged as an outline of the normal activities and sequence of events involved in eating out, and would mention such things as menus, wine lists, the various social conventions to be observed, and so on. Schemata are also invoked to explain the knowledge involved in performing a skilled task, such as driving a car, or playing chess. The learning of such tasks is seen as a matter of creating and refining a system of rules and procedures.

Propositional accounts of knowledge are inflexible because they are not sensitive to context; each different sort of situation needs its own schema, and schemata are too structured to be able to cope with the variations and novelties that exist in the environment. The result of these problems is that a multitude of schemata are necessary and this begins to look unwieldy and inefficient; it seems implausible that we should be walking encyclopaedias.

Connectionist networks account for these sorts of cognitive abilities by rejecting the notion that there are any explicit schemata. This respects Ryle's claim that propositional knowledge (knowing that p) is dependent upon knowing how to carry out certain activities, inverting the form of explanation to which conventional AI is committed. Ryle, in behaviouristic mode, analysed knowing how in terms of dispositions to produce a skilled behaviour in the relevant circumstances. However, connectionism offers a plausible way of explaining knowing how that does involve talk of internal processes. Thus McClelland and Rumelhart *et al.* (1986) argue that:

> schemata are not "things". There is no representational object which is a schema. Rather, schemata emerge at the moment they are needed from the interaction of large numbers of much simpler elements all working in concert with one another.[17]

Knowledge (or more parsimoniously, information) is not stored explicitly in a connectionist network, but implicitly as alterations in the connection weights, and is constituted by a pattern of activation across the units of the network. When required this pattern can be recreated by an appropriate stimulus, because of the changes made to the connection weights. The connections between a relatively small number of units are capable of accurately storing many different patterns. These features endow networks with a capacity (which is also displayed by humans) known as content-addressable memory, because a memory can be accessed by a partial description of its content. Such systems are tolerant of errors; access can still be gained even when some of the content cues are wrong, which allows networks to cope in informationally harsh environments. This contrasts with conventional AI where information is stored explicitly in propositional format at a particular *location*, and in order to retrieve it the address must be totally correct.

Given a partial description as a cue, a network can be thought of as conducting a best-fit search for the complete pattern. This capacity is explained by what is called soft constraints

[17] McClelland, J. L., Rumelhart, D. E., and the PDP Research Group. 1986. p.20.

satisfaction. Each unit receives excitation and inhibition from a large number of other units, all of which can be interpreted as placing multiple constraints on the units behaviour. The unit satisfies these constraints by taking an activation level which is a compromise over all its activations according to its activation function. Thus if the input pattern is largely inhibitory the unit will not fire, so the constraints imposed on the unit are soft, because none has overriding control. When an incomplete pattern is presented to a network it will satisfy some constraints, but not others. The pattern into which the network settles will be the one which produces the best compromise between these many micro-constraints, and in most cases this is the correct pattern.

An advantage of this system is that context can be encoded as part of the pattern of activation, and thus as part of the memory, and so the context within which recall takes place will effect the pattern of activation produced. The crucial point is that this pattern does *not* remain constant across instances, this is what is meant by semantic opacity. The pattern of activity which represents something in the environment is sensitive to context:

> the context alters the internal structure of the symbol: the activities of the sub-conceptual units that comprise the symbol - its subsymbols - change across contexts. In the symbolic paradigm the context of a symbol is manifest *around* it and consists of *other symbols*; in the subsymbolic paradigm the context of a symbol is manifest *inside* it, and consists of subsymbols.[18]

Individual units can be considered to represent microfeatures (subsymbols) of the environment. Microfeatures are artificial constructs useful in interpreting the behaviour of networks. They are features which are more fine-grained than those usually discussed within the semantic domain. At the semantic level of plants a flower is a microfeature, but a flower could also be dismantled into other microfeatures such as petals, stamen, and pollen. Thus there are many possible levels at which networks can extract regularities and patterns from the environment. In some networks individual units do represent particular microfeatures, especially the input and output units. For example, it might be decided that the firing of an input unit represents a given microfeature. However, in networks with hidden units the ascription of microfeatures can only treated as a useful interpretative tool. This is because these units are responding to the activity of many units in the previous layer, and thus even if these units are semantically interpretable in terms of microfeatures, the hidden units will not be. A given hidden unit may fire more to a particular environmental regularity, but it will not respond exclusively to it, because it will be involved in the encoding of many different regularities. This is a result of the holistic and distributed nature of the information storage in connectionist networks. The way in which networks come to represent the world is discussed more fully in section 4.

Another consequence of distributed representation which matches human cognition is the ability to generalise. As noted above, a new piece of information is stored by slight modifications in the connection weights of the network. If two very different patterns are stored on a network there will be virtually no interference between the them. If, however, a number of similar patterns are stored then the changes in connection weights brought about by each pattern will have a tendency to interfere with one another. This interference, far from being a problem, allows networks to form prototypes, through the cumulative effect of the many small changes in connection weights. Despite this effect networks normally retain the ability to recall specific examples of a category, and it has even been demonstrated that several prototypes can be extracted and stored by a single network.

This property of networks provides a highly plausible substrate for Wittgenstein's account of learning. Learning, for Wittgenstein, was a matter of presenting examples. The pupils grasping of the concept, or rule, being determined by their brute ability to go on and use it successfully in future situations:

> what has the expression of a rule - say a sign-post - got to do with my actions? What sort of

[18] Smolensky, P. 1988. p.17.

> connection is there here? - Well, perhaps this one: I have been trained to react to this sign in a particular way, and now I do so react to it.[19]

No structure of internal symbols could count as the understanding, because every symbolic representation can be interpreted in many different ways; and in turn each interpretation, if it is symbolic, will also be in need of further interpretation *ad infinitum*. The individual units of a network are semantically indeterminate, it is only as a whole that the network can be considered to process, store, recall, and generalise. Thus while connectionism supplies an account of internal mechanisms, it does this in a manner sympathetic to Wittgenstein's philosophy.

Another useful aspect of connectionist networks is known as *graceful degradation*, by which it is meant that the performance declines gradually when the network is put under pressure, or is damaged. In contrast conventional AI systems tend to crash if they are damaged, because they consist of production rules, which take the form *If A then B*. These are hard constraints because they show no sensitivity to other events occurring simultaneously in the environment. If a production rule is no longer present then the aspect of cognition that it accounted for can no longer be produced. On the whole human pathology displays graceful degradation rather than the lose of discrete cognitive capacities, often coping remarkably well with gross physical insults.

As briefly outlined above, connectionism appears to explain many aspects of human cognition in an elegant and unforced manner, whereas conventional models appear brittle and over-engineered. In part this might be expected because of the fact that connectionism is neurally inspired. This, however, is merely one aspect of connectionism's appeal from the naturalistic perspective, the brain is *only* the inspiration, and is not seen as the precise architecture to which it must eventually (and necessarily) aspire. A deeper reason for accepting connectionism as a correct account of mind is that its distributed form of representation respects evolutionary considerations in the development of cognition.

Human cognition is a product of evolution, as much as any part of the human anatomy, and this places certain constraints on its developmental history. This means that it is necessary to consider both the sorts of problems that would be faced by a primitive organism, and the ways open to evolution to solve these problems. Clark lists the most obvious adaptive features in cognition as:

> real-time sensory processing, integration of various input and output modalities, capacity to cope with degenerate and inconsistent data, and flexible deployment of available cognitive resources.[20]

Such abilities are needed in a harsh, competitive environment. In a world full of predators the fast processing of information is more important than getting it right every time. It is better to be a bit jumpy than very dead, and as the environment is an informationally messy place, this requires the ability to deal with degraded and ambiguous data. A robust cognitive constitution is also needed, which will not crash when overloaded or damaged. Such basic organisms will also be primarily concerned with *sensory* information and a repertoire of behavioural responses. This begins to sound very much like an advertisement for connectionism, in that parallel processing is fast, robust (due to satisfaction of soft constraints and content addressability), degrades gracefully and was originally designed for pattern recognition.

An important principle of evolution is that it is gradual, complex organisms are formed by the slow accumulation of small changes. At each stage in this progression the only important interaction is that between the organism and the environment. Thus there is no mechanism whereby a maladaptive intermediate can be maintained in order to produce a well adapted organism in the future. Each stage must be adaptive enough to survive on its own merits within its ecological niche. In addition to this, evolutionary solutions are dictated by the materials at hand, the more complex an organism becomes the greater the role its history plays in future development. For at any point in an

[19] Wittgenstein, L. 1953. §198.
[20] Clark, A. 1989. p.62.

organisms evolution, past design solutions limit the "problem space" within which evolution can work, and this tends to mean that evolutionary solutions tend to be cobbled together, rather than elegantly engineered. Evolution is a *satisficer* - Herbert Simon's term (1957) - rather than an optimiser. So we end up with lungs evolved from a fishes air bladder, dolphin flippers made out of feet and panda's thumbs fashioned from wrist bones.[21] As Clark remarks:

> the implications for what is fundamentally a *design-orientated* cognitive science [conventional AI] may be profound. For why suppose that cognitive adaptions are exempt from the same constraints? To put the point starkly, why suppose that our means of, say, playing chess is not fundamentally informed by the natural constraint of building a chess-playing capacity out of cognitive components designed for spotting predators?[22]

Philosophers have gone astray in exactly the same way as cognitive scientists. They have made an implicit adaptionist assumption about the features of human cognition that *seem* the most salient and essential. This mistake is illustrated by Gould and Lewontin (1978). They take as an example the spandrels of Saint Marks Cathedral in Venice. A spandrel is the triangular space between the curves of adjacent arches, and as such is an architectural by-product of them. This shape does not usually lend itself to decorative exploitation, but in Saint Mark's they are exploited well to create a rather prominent artistic feature. Thus one might be tempted to explain the overall structure in terms of the need to create the spandrels, but this is to entirely misunderstand the course of events that led the artist to paint the spandrels. In the case of cognition the spandrels might be logical reasoning, language and planning and the arches pattern recognition, sensorimotor co-ordination and selective attention.

Language is the central cause of this misconception. Simply by virtue of the fact that we *use* language to pick out and denote mental states, this distorts our perspective, causing us to attempt to explain cognition in linguistic terms. It is probable that language is produced by symbolic mechanisms implemented on the parallel wetware of the brain in some as yet understood way. However, I interpret the above arguments as strong evidence against attempting to use the categories of language to explain the majority of what goes on in the brain - this is just as wrong as trying to explain arches in terms of spandrels. The upshot of this is that mental states are not semantically transparent; there is no language of thought. We need to approach the mental in a new way, that explains how we come to talk about beliefs and desires (and other "mental states") without them being reducible to discrete entities in the head.

[21] Gould, S. J. 1980.
[22] Clark, A. 1989. p.72.

4. Distributing Belief

The status of folk psychology and its postulates (beliefs, desires, hopes and fears) has been problematic for philosophers. It appears that we cannot manage without this conceptual framework, yet exactly what form, if any, it takes in the head is a troublesome question.

As discussed above, there are those under the influence of the discrete entity picture, who treat cognitive processing as semantically transparent; there is a symbolic entity in the brain that co-varies with the belief. This form of realism about beliefs raises more questions than it solves. What is the role of language in belief, are beliefs sentences in the head, either in English or mentalese? If they are, is it possible for animals and preverbal humans to have beliefs? (Would we not want to say that a dog believes that beef steaks are tasty?) Even if beliefs really are sentential in form, this still does not adequately explain how beliefs have content. If a belief is a sentence it is still in need of interpretation, and as Wittgenstein pointed out:

> . . . any interpretation still hangs in the air along with what it interprets, and cannot give it any support. Interpretations by themselves do not determine meaning.[23]

Any sentence can be interpreted in many ways, and thus if one relies on more sentences to remove the ambiguity this would lead to an infinite regress (as discussed in section 3). Attempts to solve the problem of reference within the language of thought framework, such as causal or indicator theories, have all failed. The reasons for this will be discussed below.

Realism also seems to imply a proliferation of beliefs, for there is nothing in the folk notion, or even in the philosophical notion, of belief which could limit the number we ascribe to someone. As Dennett puts it:

> . . . common intuition does not give us a stable answer to such puzzles as whether the belief that 3 is greater than 2 is none other than the belief that 2 is less than 3.[24]

Beliefs are invoked in order to explain people's behaviour. If they are sentential, then it would be possible to attribute a vast array of different beliefs, and other propositional attitudes. This is because of the inherent ambiguity that lies between propositional attitudes and the behaviour they are supposed to explain. What happened in peoples heads all around the world when they heard that President Kennedy had been assassinated? Surely we would want to say that all over the world people were acquiring *the same belief*. Yet this single belief could be expressed by many different linguistic, or symbolic, formulations. The only way that these various formulations could be grouped together would be as tokens of the same belief, the very thing that they are supposed to explain.

The problem here originates in the assumption of semantic transparency. However, denying semantic transparency makes the ontological status of propositional attitudes problematic. P. M. Churchland (1981) has argued that folk psychology is a defunct *theory* about human behaviour, and that it should be eliminated in favour of a more neurophysiologically accurate theory:

> Not only is folk psychology a theory, it is so *obviously* a theory that it must be held a major mystery why it has taken until the last half of the twentieth century for philosophers to realise it. The structural features of folk psychology parallel perfectly those of mathematical physics; the only difference lies in the respective domain of abstract entities they exploit - numbers in the case of physics, and propositions in the case of psychology.[25]

Such a view fails to appreciate the possible relationships that can exist between different types of explanation. First it assumes that folk psychology is a theory on all fours with other scientific theories. This is to be seduced by language, for my point will be that folk psychology cannot be set out, or applied, as a series of laws, even if 'proper' scientific theories can be fully explicated in such a fashion. Secondly Churchland demands a smooth reduction of folk psychology to

[23] Wittgenstein, L. 1953. §198.
[24] Dennett, D. C. 1987. p.55.
[25] Churchland, P. M. 1981. p.71.

neurophysiology, which is an overly high standard for theoretical vindication. The notion of functional utility applied to designed objects is clearly dependent upon arrangements of physical particles, but cannot be reduced to any notion defined in purely physical terms. However, this does not indicate that the notion of functional utility needs to be abandoned. The proper response, then, is to keep the discussion of folk psychology and its physical substrate separate. Dennett argues that we should do this by assuming the intentional stance. This involves treating people as idealised rational systems, where:

> . . . intentional systems theory is envisaged as a close kin of, and overlapping with, such already existing disciplines as decision theory and game theory, which are similarly abstract, normative and couched in intentional language. It borrows the ordinary terms "belief" and "desire" but gives them a technical meaning within the theory. It is a sort of holistic logical behaviourism because it deals with the prediction and explanation from belief-desire profiles of the actions of whole systems (either alone in environments or in interaction with other intentional systems), but it treats the individual realisations of the systems as black boxes. The subject of all the intentional attributions is the whole system rather than any of its parts, and individual beliefs and desires are not attributable in isolation, independently of other belief and desire attributions.[26]

Thus one attributes the beliefs and desires that an individual *ought* to have, given their basic needs, and then one decides what they *ought* to do to accomplish their desires. This gives folk psychology the same status as centres of gravity: they are both constructs useful in calculation, but have no reality beyond this. As a complete account of the mental the intentional stance is untenable, because it inherits the severe defects of behaviourism. However, it points the way to a proper construal of folk psychology through its appreciation that intentionality and rationality should be understood at an abstract level. The problem for a non-linguistic account of human cognition is to account for our inferential and logical abilities. Once it is realised that such notions are abstractions, and thus do not form part of our basic mental furniture, this problem begins to look less serious.

Dennett argues that our notion of rationality derives from evolutionary concerns, because cognition evolved through tracking salient and important items in the environment in the right sort of way:

> Treating each other as intentional systems works (to the extent that it does) because we really are well designed by evolution and hence approximate to the ideal version of ourselves exploited to yield the predictions. But not only does evolution not guarantee that we will always do what is rational; it guarantees that we won't. If we are designed by evolution, then we are almost certainly nothing more than a bag of tricks, patched together by a *satisficing* Nature, and no better than our ancestors had to be to get by.[27]

Thus philosophers have misunderstood the nature of the link between the philosophical notion of rationality, or logic, and the folk notion. They have attempted to distil the folk concept down to its essence; but it cannot be treated in this way, as Wittgenstein has argued:

> . . . logic does not treat of language - or of thought - in the sense in which a natural science treats of a natural phenomenon, and the most that can be said is that we *construct* ideal languages. But here the word 'ideal' is liable to mislead, for it sounds as if these languages were better, more perfect, than our everyday language; and as if it took the logician to show people at last what a correct sentence looked like.[28]

The interpretation of language and behaviour, with which practical folk psychology is concerned, does not rely on the use of an explicit (or implicit) system of rules. Rather it should be seen on the model of pattern recognition described in section 3: through being presented with series of examples we learn how to employ a concept, or attribute a belief. Knowledge about the consequences of attributing a particular belief is constituted by a set of expectations and behavioural dispositions. In

[26] Ibid. p.58.

[27] Ibid. p.51.

[28] Wittgenstein, L. 1953. §81.

this way knowledge of what to do in a given situation arises out of the context of that situation, through its associations and similarities with previously experienced situations. Folk psychology, as used by ordinary people, does to a certain extent deal with idealised, abstract entities because it involves talk of beliefs and desires. However, the idealised rationality which forms the basis of the intentional stance is a more idealised propositional description of the patterns and relationships of behaviour and judgements produced by ordinary folk psychologists.

It also follows from the fact that human cognition is an evolutionary product that we only approximate to the intentional stance, it applies in most, but not all situations. Logic does not reflect the world - it is an abstraction from it. Churchland's eliminativism is brought about by two errors. Firstly, he fails to make a distinction between ordinary folk psychology and the intentional stance, and thus he argues that folk psychology is a pseudo-scientific theory. Secondly he argues that folk psychology's failure to reduce to neurophysiology counts against it, but this is because he fails to see that its "abstract entities" are part of an idealised conception; folk psychology and the intentional stance make no claims about implementation.

Thus the intentional and implementational approaches need not be seen as conflicting *if* they are interpreted in the correct way. In this spirit Clark suggests that there should be a distinction between two kinds of cognitive science:

> *Descriptive cognitive science* attempts to give a formal theory or model of the structure of the abstract domain of thoughts, using the computer program as a tool or medium.
>
> *Causal cognitive science* attempts to give an account of the inner computational causes of the intelligent behaviours that form the basis for the ascription of thoughts.[29]

These can be equated with what Dennett terms Intentional Systems Theory and Sub-personal Cognitive Psychology. The trouble for traditional AI results from the fact that it aspires to causal cognitive science, whereas it should really only concern itself with descriptive cognitive science. Connectionism, on the other hand, can be considered as causal cognitive *science par excellence*. Connectionism gives an account of how it comes about that our cognitive systems display intentional characteristics without actually embodying them in a formal and transparent manner.

Conventional AI can be seen as modelling the intentional stance, attempting to formulate rules which capture the structure of human behaviour. So conventional AI computer models of the mind have the same sort of status as computer models of the weather, they give a description of the patterns exhibited by the actual phenomenon. Thus they succumb to Searle's arguments about causal inefficacy:

> We do not suppose that because a weather program simulates a hurricane, that the causal explanation of the behaviour of the hurricane is provided by the program. So why should we make an exception to these principles where unknown brain processes are concerned?[30]

Searle's point is that computers deal only with syntax, they are *syntactic* engines, whilst we are *semantic* engines. No consideration is given to the reference of a symbol in processing, rules are applied in virtue of the symbols syntactic form only. The symbols of conventional AI are treated as arbitrary tokens, and so there is no link between the properties of the symbol and its reference in the world. There is no mechanism whereby the symbol can get to be about an object in the world, it "passes over" the world in the terminology of Heidegger. As a consequence of this the symbols are context free: a symbol has the same meaning (for an external observer) regardless of context. However, with intentional representations the intended referent of a term varies with context. Thus in one situation the word "bat" might refer to a nocturnal animal, and in another it might refer to an implement used to strike a ball. As described in section 3, conventional AI attempts to avoid this by postulating different symbols for each meaning, but this still leaves the problem of ascertaining the correct symbol to use in a given situation. According to Searle such systems will always be unable to

[29] Ibid. p.153.
[30] Searle, J. R. 1992. p.218.

derive semantics from syntax.

> What matters about brain operation is not the formal shadow cast by the sequence of synapses but rather the actual properties of sequences.[31]

Hence there is no way that computer models can exhibit intentionality, because they do not have the right kind of causal properties; they merely describe the system according to an externally specified - and thus observer dependent - pattern.[32] Searle's worries are (to a limited extent) legitimate, but his appeal to causal properties as the haven for intentionality is ambiguous. The source of his concern is in the way conventional computer models fail to capture the immense structural variability and complexity with which the brain is endowed. This gives rise to his infamous Chinese room thought experiment, but this only attacks models which are semantically transparent. The thought experiment can be bypassed if we angle our formal description at a lower, more microstructural level. The connectionist approach respects the complexity of human behaviour, and allows a causal account of the processes underlying intentionality and semantics to be given. As Dennett points out, Searle's mistake is his desire for an explanation of *intrinsic* intentionality:

> How could any entity get the semantics of a system from nothing but its syntax? It couldn't. The syntax of a system doesn't determine its semantics. By what alchemy, then, does the brain extract semantically reliable results from syntactically driven operations? It cannot be designed to do an impossible task, but it could be designed to *approximate* the impossible task, to *mimic* the behaviour of the impossible object (the semantic engine) by capitalising on close (close enough) fortuitous correspondences between structural regularities - of the environment and of its own internal states and operations - and semantic types.[33]

Connectionism has the advantage of being conceptually continuous with tasks which are primitive in evolutionary terms. Information enters the system through its sensory transducers and passes smoothly through the system, there is no point where it becomes symbolised, producing a semantic gap between the organism and its environment. Connectionist systems react to patterns and regularities in the environment in a way that is not arbitrary. Networks learn by altering the weights of their connections, and this is done in such a way that the weights come to reflect the patterns in the environment. The *goal* of the learning procedure is to maximise the fit between the environment and the networks pattern of connection weights, thus these patterns can be seen as *representing* the external world for the system. A particular network cannot function outside the environmental niche in which it was trained, and so can be thought of as being enmeshed in the world in a non-arbitrary fashion; its states are *about* the world. This is reflected in the way networks deal with context, an important facet of intensional logic. Networks operate by reacting to patterns, thus the context within which an object is presented constitutes a part of this pattern, and thus will affect the network's processing. An interesting corollary of this is that the system's *internal* environment has the same capacity to affect processing as the external environment, as the system makes no such distinction between internal and external information. This suggests a possible explanation of how intentional states can be about non-existent objects or states of affairs: their content is provided by the internal (rather than external) environment.

Dennett argues that the brain only approximates the impossible object - the semantic engine - but these properties of connectionist networks show how the brain might manage to *be* a semantic engine. Connectionism demonstrates how the brain capitalises on the "fortuitous correspondences between structural regularities - of the environment and of its own internal states and operations - and semantic types". Searle demands that humans must have *real* intentionality, not the *as-if* intentionality of the intentional stance. I have tried to explain, with much hand waving, how connectionism might offer a naturalistic account of intentionality, which is not the ersatz intentionality of the intentional stance, nor an occult property of human minds. Philosophers should be wary of any

[31] Searle, J. R. 1980. p.300.
[32] This also explains why the intentional stance cannot be a complete account of the mental.
[33] Dennett, D. C. 1987. p.61.

argument that places only humans within a charmed circle. The only way in which humans are special is through the level of complexity manifested by our neurophysiology, of which language is the most important by-product. It has been assumed that linguistic representation provides the basis for intentionality. In contrast, I believe that it is the non-linguistic capacities of distributed networks which underwrite intentionality. Ironically, Searle has recognised the error in the traditional conception of mental representation:

> I now think the real mistake was to suppose that there is an inventory of mental states, some conscious, some unconscious. Both language and culture tend to force this picture on us. We think of memory as a storehouse of propositions and images, as a kind of big library or filing cabinet of representations. But we should think of memory rather as a *mechanism* for generating current performance, including conscious thoughts and actions, based on past experience.[34]

Here Searle has recognised that one of the theoretical consequences of the discrete state picture is false, but he fails to fully appreciate the consequences of a non-linguistic foundation for intentionality. For Searle the ontology of unconscious mental states is totally neurophysiological. At some times the complex physical events in the brain cause conscious states, with all their attendant (and problematic) properties. Other than these conscious mental states there is nothing but neurophysiology. Unconscious mental states are to be understood in terms of capacities of the brain to generate conscious mental (and intrinsically intentional) states. Further, these intentional states require a set of background capacities to determine their conditions of satisfaction (which Searle calls “the Background”). Some of these background capacities will be capable of generating other intentional mental states, but others will not. Thus Searle tries to provide a basis for intentionality by positing a non-intentional grounding. He wants to be able to defend intrinsic intentionality from sceptical attacks, such as the indeterminacy of translation. According to Searle’s theory, when one is faced with an intentional state (with all the bells and whistles of consciousness) its true meaning can be settled by appeal to the background capacities. We may not be able to discern them, but there *will be* a matter of fact in there somewhere. Quine gives, as an example, a community of remote islanders who use the word “gavagai” when they see a rabbit. Indeterminacy of translation means that we would not be able to ascertain whether by “gavagai” they mean “rabbit” or “undetached rabbit part” or “rabbit time slice” and so on. However, Searle would argue that there is a matter of fact, determined by their non-intentional non-linguistic background capacities.

Unfortunately this defence is inadequate to the task. Searle is extremely vague in articulating what sort of things these background capacities actually are, although he does not see this as a problem:

> It sometimes seems as if the Background cannot be represented or made fully explicit. But that formulation already contains a mistake. When we say that, we already have a certain model of representation and explicitness.[35]

This retort would be perfectly correct *if* Searle did not want his intentionality to be intrinsic, which in this case appears to imply that intentionality cannot be artificially produced, and that scientific and naturalistic explanations will always be deficient. *Pace* Searle *et al.* I have sketched a possible way that intentionality could be fully explained using a distributed connectionist model. A consequence of this approach - admittedly controversial - is that intentionality just is the mirroring of the environment in a certain way. Beliefs and desires are attributed against the background of patterns of behaviour. These patterns are real, and it is through their existence that linguistic attributions gain utility. This is just as applicable in ones own case as it is for others, except that in self-ascription one has access to the internal cognitive environment responsible for the patterns in behaviour and dispositions. There is a mass of neural activity, involved in producing internal mediating responses to environmental events (both internal and external), which generates content. We explain this to ourselves and others by

[34] Searle, J. R. 1992. p.187.
[35] Ibid. p.193.

producing linguistic descriptions, both vocally and subvocally. These linguistic descriptions, however, will always be an *interpretation* of the distributed brain processes upon which they depend. While brain processes are open to physical investigation, no matter how much information we gain about them, this will always underdetermine a particular belief attribution. To speak of the Background as a set of non-linguistic (cognitive) capacities is not to settle the matter, it merely begs the question. To call the mass of non-linguistic processing "the Background" is permissible, but to argue that this can determine linguistic attributions is to be in error. Searle is right to argue that it is a mistake "to suppose that there is an inventory of mental states, some conscious, some unconscious," but he has not recognised the full consequences of this admission.

Cognitive processing does not involve individual, isolable states and processes, but they are not necessary for intentionality if it is understood correctly. The content of propositional attitudes is parasitic on the intentionality produced by the underlying distributed neural processing, as that has been explained in this section. The limited indeterminacy of propositional attitude attribution is the result of the interface between linguistic and non-linguistic systems of representation. The non-linguistic system cannot be given a linguistic characterisation because it is not semantically transparent. Thus the intentional stance is necessarily an "abstract, idealising, holistic" process, because it uses a linguistic approach to capture the complex behavioural patterns produced by a non-linguistic system which is itself the product of millions of years of evolution. This history entails that the brain will have developed and succeeded because it bought survival value in coping with a complex and hostile environment, not because it was destined to be the vessel for a divine essence. Does all of this mean that propositional attitudes do not exist? My response is Wittgensteinian:

> Say what you choose, so long as it does not prevent you from seeing the facts. (And when you see them there is a good deal that you will not say.)[36]

Beliefs *are* real - but not in the same way as physical objects - they belong to a different existential category. Belief-talk is inescapable, but as long as care is taken to avoid metaphysical confusion, it is also harmless.

[36] Wittgenstein, L. 1953. §79.

5. The Connectionist Beetle in the Box

The intended target of Wittgenstein's private language argument was sensation, or phenomenal consciousness. In the two previous sections I have attempted to outline an approach to the mental which avoids the problems expressed by Wittgenstein. The last bastion of these misconceptions is phenomenal consciousness, and if the distributed model of consciousness is to be judged successful it must dispel the mists of confusion surrounding this redoubt. This will involve following a course between Wittgenstein and the massed defenders of qualia. By neither accepting qualia, nor denying them, but pointing out the mistakes in looking at the debate in this way, it will be possible to settle the problems and paradoxes surrounding qualia.

In order to gain a proper understanding of phenomenal consciousness it is vital to remember that it is a product of evolution. As with other aspects of cognition, described in section 3, this has important implications. If we take vision as a paradigm case, it is a form of representation that has evolved to provide information about salient features of the environment to the system. This explains the particular "shape" of visual quality space, why some colours are more prominent than others, because of their significance in terms of survival value. It also explains why the visual system fails to react to, or represent, other features; vision has not evolved in order to satisfy our every epistemic whim. Vision is constituted by content-laden activity passing smoothly through the system in a parallel and distributed manner, in accordance with Dennett's multiple drafts model.

This account of phenomenal consciousness does not mix easily with contemporary philosophical debate, which revolves around the question of qualia and behavioural dispositions. Qualia are considered to be individual *states*, with a variety of functional inputs, or causes, and outputs, or behaviour. This gives rise to a debate about whether a quale is something over and above its functional relations, something which can only be grasped from a first person perspective. However, as I have shown in sections 3 and 4, the notion of isolable states is inappropriate, and does not fit with the way the brain works (it can also be seen as implicitly presupposing a Cartesian Theatre, with all its inherent difficulties, as discussed in section 2). Rather there are networks, with input sides and output sides. One can talk about events closer to one side of the network than the other, but the only way to interpret the network is to look at the patterns in its behaviour as a complete system. Representations are distributed throughout the network, so that it is impossible to single out a particular part and give it a definite functional role analysis. This view of cognition can be employed at various levels of abstraction, from the fine-grained neurophysiological level to the coarse-grained information processing level. The latter is more or less captured by the multiple drafts model: patterns of visual information flow from the retina to the primary visual cortex and then to a number of processing systems. In these systems content-fixations are made which affect many other systems, some of which result in behaviour, or speech, whether in the head or out there in the world. This activity occurs in parallel, pandemonium style, with patterns of information flowing and being acted upon in many different ways in many different parts of the brain. This approach respects what is known about brain physiology, as philosophy should, but it also has a direct influence on the philosophy of phenomenal consciousness.

One might say that it distributes the quale throughout a network which forms part of the overall system (where this overall system is conceived as many different sub-networks, with their own sub-tasks, all linked in a parallel fashion to form the complex whole). There is no state that is the quale, but the phenomenal aspect (or sensation) is constituted by the activity occurring in the network. I am using the term "sensation" here to refer to the component of cognitive processing that gives rise to the belief that there is a quale inside the head. Dennett is an eliminativist about qualia, for him there is only *the text* of heterophenomenology and content-fixations. In contrast, I propose to conceptually redefine phenomenal consciousness in a way that avoids qualia induced problems, while maintaining that there is something there to talk about. For me the content-fixations produce a form of representation which is more than just representation *that* there is an *x* in front of me. This involves a blurred boundaries approach, according to which sensations amount to the activity of distributed parallel processing units. The question of exactly which units form the substrate for conscious experience is thus one that does not always admit of a definite answer. This is due, in part, to the fact that we have to refer to sensations using language. If I am having an experience which I describe as "seeing red" this presupposes certain conceptual distinctions and categorisations on my part; and it is here that Dennett's multiple drafts model is apt at accounting for verbal reports about phenomenal experience. Furthermore, the act of discriminating 'conscious states' carves off a part of the continuous flow of processing. This point can be brought out using a Wittgensteinian metaphor:

> . . . imagine having to sketch a sharply defined picture 'corresponding' to a blurred one. In the latter there is a blurred red rectangle: for it you put down a sharply defined one. Of course - several such sharply defined rectangles can be drawn to correspond to the indefinite one. - But if the colours in the original merge without a hint of any outline won't it become a hopeless task to draw a sharp picture corresponding to the blurred one?[37]

Attending to an aspect of visual experience (which one does when 'pointing' internally to the quale, remarking "This!") is like trying to sketch a sharply defined picture from a blurred one. The processing at the centre of the patch of red will be contributing to the sensation, but as one moves nearer the periphery the role of the processing units becomes less clear, until eventually one reaches the area on the outside of blurred rectangle, where the units are definitely not contributing to the experience. This metaphor should not be taken too literally, however. For it may be the case that a number of distal areas contribute to the sensation, the point of the metaphor is that as representations are distributed it is difficult to ascertain whether any particular group of units is involved and there is just no fact of the matter in some instances.

This indeterminacy stems from the fact that the activity of individual units is not semantically interpretable. Content is distributed, and there is no method for determining the limits of this distribution within the system. This is largely an empirical matter, but it seems certain that the whole brain will not be involved in a particular visual experience. My hunch is that several different sub-networks will be involved, and that these networks will differ with context; where this could include the type of phenomenal consciousness (visual, aural etc.), the task being carried out (object recognition, motor control etc.) and feedback from other aspects of processing (such as sensorimotor feedback).[38] Due to the parallel, multi-track, nature of processing functional inputs and outputs are distributed throughout the system, or perhaps it would be more accurate to say that the distinction between inputs and outputs loses its utility. The outputs to some subsystems act as inputs to others, so the linearity of the traditional view is compromised. This view has it that inputs are the information on the way in (to the Cartesian Theatre, or the quale, or whatever) and that outputs occur after this, on the way out to behaviour. If there is no clear point for them to meet up then the notions become relativised. What constitutes an input depends on the part of the system that you are concentrating on, and on the task the system is carrying out. There are only micro-inputs and outputs (micro-dispositions): the excitations that units receive and give out, this is all that occurs in the brain. The notion of a quale is rejected and replaced by an indeterminate cloud of micro-dispositions with which the sensation is identified. A consequence of this is that there is no change in the sensation without a change in micro-dispositions, and vice versa. Thus this theory amounts to a sophisticated version of functionalism, which is easier to swallow because it shows more respect to the complexities of consciousness and cognition. In addition it illustrates the conceptual problems manifested in the many thought experiments, or intuition pumps, concerned with qualia.

Firstly it accounts for Dennett's Chase and Sanborn intuition pump in a way that upholds his

[37] Wittgenstein, L. 1953. §77.

[38] Zeki, S. 1977, has conducted work on the visual system of the rhesus monkey which suggests that there are richly interconnected specialised regions for analysis of various aspects of vision, such as colour, form, and movement.

verificationism without denying sensation.[39] Chase and Sanborn are coffee tasters who have both come to dislike the taste of Maxwell House, where once they enjoyed it. Chase believes that the coffee tastes the same as it always did, but that his taste preferences have changed, he no longer likes *that taste.* Sanborn, on the other hand, believes that the taste of Maxwell House has changed for him, while his preferences have remained constant. Dennett argues that there is no way to differentiate between these two cases. Any given case can be described in these two ways, as quale constant with preference (or memory) shift, or quale shift with preferences (or memory) constant, or a point intermediate between these two extremes. This last option is important, because it allows Dennett to fend off psychological and neurophysiological evidence, for it seems obvious that it would be relevant in dealing with the extreme cases:

> Thus if Chase is unable to reidentify coffees, teas and wines in blind tastings in which only minutes intervene between first and second sips, his claim to know that Maxwell House tastes just the same . . . will be seriously undercut.[40]

If, however, there is always the possibility of a composite error, the empirical evidence will never be fine-grained enough to determine the matter. This creates a scenario in which no evidence *whatsoever* could decide, neither objective nor subjective, because the experiential reports of Chase and Sanborn are fallible on this interpretation of qualia. They will not be able to tell what has shifted and what has remained constant, because there could always be intermediate cases which would be phenomenally indistinguishable. As Dennett has remarked:

> The standard rebuttal to this verificationist assertion is that I am prejudging the course of science; how do I know that new discoveries in neuroscience won't *reveal* new grounds for making the distinction? The reply - not often heard these days - is straightforward: about some concepts (not all, but some) we can be sure we know enough to know that *whatever* came along in the way of new science, it wouldn't open up this sort of possibility. Consider, for instance, the hypothesis that the universe is right-side-up, and its denial, the hypothesis that the universe is upside-down. Are these hypotheses in good standing? Is there, or might there be, a fact of the matter here? Is it a verificationist sin to opine that no matter what revolutions in cosmology are in the offing, they won't turn that "dispute" into an empirical fact of the matter that gets settled?[41]

To talk of qualia and dispositional reactions to them is as much a nonsense as to argue over whether the universe is right-side-up, and verificationism in such situations is wholly warranted. The 'object and judgement' paradigm is wrong, and the distributed model of consciousness shows how it is wrong: because the sensation is constituted by the micro-dispositional reactions. Thus it is not possible for the phenomenal aspect of a sensation to change whilst dispositions and preferences remain the same. By making this identification it has been my intention to demonstrate that there is more going on than Dennett has admitted. While there are no qualia, there is still phenomenal experience.

In this fashion the distributed model gives an explanation of the problems surrounding the inverted spectrum intuition pump. In order to accomplish this, it is necessary to make some speculative empirical assumptions which may be falsified by future research. The source of all the trouble is the ease with which we can think in terms of colours. When we do this we imagine little colour patches, and this encourages us to imagine slipping one colour patch out of the functional web, and putting in another one. This is also encouraged by the apparently easy neural switcheroo between the three different types of colour receptors in the eye. We can imagine the nerves for the receptors responsive to green being switched with those from the receptors responsive to red. Yet empirical findings suggest that the experience of colour is determined by processes much more complex than simply the wavelength of the light falling on the retina. The colours we see are

[39] Dennett, D. C. 1988.
[40] Ibid. p.529.
[41] Dennett, D. C. 1991. p.462.

probably determined by the relations and patterns in the visual information we receive, and by expectations about this information.[42] They are thus a composite of a complicated mass of processing. If we did carry out a switcheroo on an individuals early visual system their experience would certainly be very strange, but it would not be characterised in the way that philosophers have assumed. The individuals experience would be determined by the interactions of their various sub-networks, the product of which would probably not be captured easily by language, and this is all that can be said. This can be illustrated by experiments in which people wear goggles that invert the visual field. Given time the subjects become adapted to the goggles, until they can function in a totally normal way, but to ask if they do this by turning their experiential world the right way up again, or by getting used to an inverted world is to misunderstand the phenomenon. As they adapt the subjects themselves dismiss such questions as the wrong sort to ask. The adaptation process involves accommodation by the alteration of a multitude of micro-dispositions. Some will be concerned with deliberate guided motor control, others with more reflexive reactions. The subject's experiential world is certainly different, but it admits of no easy characterisation through language.

What emerges from this is that phenomenal consciousness is an extremely plastic form of representation that is influenced by higher cognitive concerns. These higher cognitive concerns are formed by language and the epistemic and survival needs that it imposes on us. Language *is* a very important tool for the human mind, but this does not entail the existence of a language of thought as the fundamental form of brain representation. Rather language has a peculiar modulatory effect on what we experience. Wittgenstein and Dennett rightly point to this influence, but make too much of it, and the qualiophiles refuse to recognise the complex and fluid nature of phenomenal consciousness. As discussed in section 3 cognitive processing is semantically opaque, and thus non-linguistic, but it *has* given rise to language, which has allowed us to accomplish a wide variety of tasks, through the influence of evolutionary pressures. This indicates that the neurophysiological substrate of language will not be the same as the one underlying more primitive, but nevertheless essential, cognitive functions.

The distributed model of consciousness rejects the distinctions that have dominated the philosophy of mind. In the light of this model Wittgenstein's *Philosophical Investigations* can be seen as an attack on a semantically transparent view of the mental:

> Suppose everyone had a box with something in it: we call it a "beetle". No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle. - Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. - But suppose the word "beetle" had a use in these people's language? - If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a *something*: for the box might even be empty. - No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.
>
> That is to say: if we construe the grammar of the expression of sensation on the model of 'object and designation' the object drops out of consideration as irrelevant.[43]

Wittgenstein's beetle *is* the distributed sensation, the indeterminate collection of micro-dispositions, which neatly answers to his description. That everyone has something different in his box should come as no surprise. As explained in section 4, in the context of belief, there is no unique physical substrate underlying a particular belief. The belief is a linguistic description of semantically opaque cognitive processing, and thus only makes sense at the intentional level. This point applies equally in the case of sensation. As there is no separation of the phenomenal aspect of sensation from the indeterminate collection of micro-dispositions, the sensation can be thought of as inner without being mysterious. Phenomenal experience is not identified with behaviour, but micro-behaviour, and thus enters the language-game as the cause of observable behaviour and of behavioural dispositions. It is also unsurprising that the thing in the box is constantly changing, this is just a consequence of the

[42] See Zeki, S. 1993, for an excellent discussion of the neurophysiology and psychology of colour.

[43] Wittgenstein, L. 1953. §293.

nature of cognitive processing, which is highly malleable, constantly changing in response to the environment and task demands. It is suggestive that Wittgenstein equivocates about his claim that the beetle-in-the-box is nothing:

> "But you will surely admit that there is a difference between pain-behaviour accompanied by pain and pain-behaviour without any pain?" - Admit it? What greater difference could there be? - "And yet you again and again reach the conclusion that the sensation itself is a *nothing*." - Not at all. It is not a *something*, but not a *nothing* either! The conclusion was only that a nothing would serve just as well as a something about which nothing could be said We have only the grammar which tries to force itself on us here.[44]

Wittgenstein is right to resist the grammar as it tries to force conceptual distinctions upon us, but there is no need to deny utterly the link between the beetle-in-the-box and the language game. The link is merely much more complicated than we had imagined. Language describes the structure and patterns in linguistic and bodily behaviour, not the underlying neural processing. Yet this behaviour could not take place, and would not be conscious, without these processes. The distributed processing provides the mechanism for the many natural behaviours and dispositions upon which language is constructed; it is what makes the difference between pain-behaviour accompanied by pain and pain-behaviour without any pain. In this light one of Wittgenstein's most celebrated attacks against an internalist account of cognition can be seen as compatible with my distributed, interpretational model:

> If I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos?[45]

Connectionist networks are not exactly chaotic, but from the semantic level they *appear* to be chaotic, due to their semantic opacity. Their behaviour can only be analysed at the level of the complete system, and in terms that only apply at that level. Thus in ascribing a belief to another person I am not talking about their internal processes, but about their behaviour at the intentional level. However, those internal processes are still needed as a substrate for that belief, and thus they become involved in the language-game.

Paradoxical as it may seem to some, I believe that if he were alive today, Wittgenstein would embrace connectionism as the proper model to explain the nature of the mind and consciousness.

[44] Ibid. §304.

[45] Wittgenstein, L. 1967. §608.

Bibliography

Bechtel, W & Abrahamsen, A. 1991. *Connectionism and the Mind*. Oxford: Blackwell.

Block, N. 1994. “Consciousness.” In Guttenplan, S. (Ed.) *A Companion to the Philosophy of Mind*. Oxford: Blackwell.

Churchland, P. M. 1981. “Eliminative Materialism and the Propositional Attitudes,” *Journal of Philosophy*, **78**, pp. 67-90.

Clark, A. 1989. *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, MA: MIT Press/Bradford Books.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press/Bradford Books.

Dennett, D. C. 1988. “Quining Qualia.” In Lycan, G. (Ed.) *Mind and Cognition*. Oxford: Blackwell.

Dennett, D. C. 1991. *Consciousness Explained*. Boston, Little Brown.

Fodor, J. 1975. *The Language of Thought*. New York: Thomas Y. Crowell.

Gould, S. J. 1980. *The Panda’s Thumb*. New York: W. W. Norton & Co.

Gould, S. J. & Lewontin, R. 1978. “The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptionist Programme,” *Proceedings of the Royal Society*, **B205**, pp. 581-598.

McClelland, J. & Rumelhart, D. and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2. Cambridge: MIT Press.

Nagel, T. 1974. “What is it like to be a bat?” *Philosophical Review*, **83**, pp. 435-450.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson.

Schank, R. & Abelson, R. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Searle, J. R. 1980. “Minds, Brains and Programs,” *Behavioural and Brain Sciences*, **3**, pp. 417-424.

Searle, J. R. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press/Bradford Books.

Simon, H. 1957. *Models of Man*. New York: Wiley.

Smolensky, P. 1988. “On the Proper Treatment of Connectionism,” *Behavioural and Brain Sciences*, **11**, pp. 1-74.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Wittgenstein, L. 1958. *The Blue and Brown Books.* Oxford: Blackwell.

Wittgenstein, L. 1967. *Zettel*. Oxford: Blackwell.

Zeki, S. 1977. “Colour Coding in the Superior Temporal Sulcus of the Rhesus Monkey Visual Cortex.” *Proceedings of the Royal Society*, **B**, pp. 195-223.

Zeki, S. 1993. *A Vision of the Brain*. Oxford: Blackwell.