Virtually Explained:

Daniel Dennett's Theory of Consciousness – Explanation and Implementation

By Stephen James Edwards

962080735

Submitted in partial fulfilment of the academic requirements for the degree of Master of Arts (in Philosophy) at the University of Natal, Durban.

1st August 2003

Declaration

I, Stephen James Edwards, hereby declare that unless specifically indicted to the contrary in this text, this dissertation is all my own work and has not been submitted to any other university in full or partial fulfilment of the academic requirements for any degree or other qualification.

Signed at Durban on the first day of August 2003.

.....

Acknowledgements

I would like to thank everybody in the Philosophy Department at the University of Natal, Durban, for their help, support, and companionship throughout the writing of this dissertation. More specific thanks go out to Andries Gouws, John Collier, Stephen Cowley, and Andrew Dellis for supplying much appreciated and helpful discussion, comments, and advice along the way. I would *especially* like to thank: David Spurrett for his patience, enthusiasm, and invaluable guidance during his supervision of my efforts; Grant Blair for making this degree much more fun, and much less lonely than it could have been; and finally, Nadine Barron, and the members of my family, whose unwavering care and reassurance has helped me enormously.

Abstract

This paper is an analysis of aspects of Daniel Dennett's theory of human consciousness. For Dennett, the reasons why human consciousness is so unique among earthly creatures, and so manifestly powerful, are not to be found in the differences between our brains and those of other higher mammals, but rather in the ways in which the plasticity of our brains is harnessed by language and culture. According to Dennett, the best way to understand the enhancements and augmentations that result from enculturation is as a *von Neumannesque virtual machine implemented in the parallel-distributed processing brain.* This paper examines two questions that arise from the latter hypothesis: (1) If non-symbolic, parallel-distributed networks accomplish all the representation and computation of the brain, what kind of explanation of the functionality of that brain, can legitimately maintain descriptions of procedures that are symbolic, serial, and real? (2) What kind of structural design, training, and resultant processing dynamics could enable a (human) brain to develop a competency for symbolic, serial procedures? Through addressing these questions, I argue that Dennett's theory of consciousness is broadly correct, investigate some other theorist's ideas that are highly compatible with Dennett's work, and consider some criticisms that have been levelled against it.

Table of Contents

Two passengers leaned against the ship's rail and stared at the sea. 'There sure is a lot of water in the ocean,' said one. 'Yes,' answered his friend, 'and we've only seen the top of it.'

- George Miller on consciousness (1962).

And what are we to say of man? Is he a speck of dust crawling helplessly on a small and unimportant planet, as the astronomers see it? Or is he, as the chemists might hold, a heap of chemicals put together in some cunning way? Or, finally, is man what he appears to Hamlet, noble in reason, infinite in faculty? Is man, perhaps, all of these at once?

- Bertrand Russell (1959)

Man acts as though he were the shaper and master of language, while in fact language remains the master of man.

- Martin Heidegger (1951)

Virtually Explained:

Daniel Dennett's Theory of Consciousness – Explanation and Implementation

Section 1: Introduction

This section will begin by describing Dennett's position on human consciousness (1.1, 1.2, 1.3). Section (1.4) isolates the major themes of this paper, outlines the position I will defend, and provides an outline of how I intend to proceed.

1.1 Setting the Scene

One of the most remarkable facts about human brains is their plasticity. Although genetic evolution has given rise to this trait, cultural evolution has exploited the adaptability and programmability that it offers, allowing for a radical enhancement of our innate abilities. Six million years ago there was a primate from whom all chimpanzees, and all human beings have descended. Human brains are now nearly four times as big as chimpanzee brains which remain of a similar size to our common ancestor.[1] Crucially, the massive growth in the size of human brains was essentially complete *before* the development of a significant language and culture, and thus cannot be a response (via the Baldwin Effect) to all the complexities that advanced language and culture impart on human brains (Dennett 1991a: 190). As a result of this fact Dennett speculates that the vast enhancement of human mental powers has mostly occurred since the end of the last ice age, that at birth, our brains are (more or less) equipped with the same powers as the brains of our ancestors some 10000 years ago, and thus the amazing advancements we have made since then, are largely "due to harnessing the plasticity of that brain in radically new ways – by creating something like *software* to enhance its underlying powers" (Dennett 1991a: 190).

[1] Although the reasons for this large and speedy growth are interesting, important, and controversial, they are not part of my endeavour here – Dennett recommends William Calvin's books for further reading. Also see Terrence Deacon's (1997) book: *The Symbolic Species*.

Setting this 'software' analogy aside for the moment, it is worth considering how Dennett thinks our ancestors might have taken the first steps towards enhancing their mental abilities. Dennett suggests that we possessed an innate curiosity, and that from this we developed strategies of 'self-stimulation' or 'self-manipulation' that produced habits that could effectively enhance the powers of our brains (1991a: 209). Dennett calls these habits 'meme-effects'. As is well known: a meme is a term originally coined by Richard Dawkins (1976) and basically refers to any cultural artefact. Dennett suggests that meme-effects (and consequently our 'neural weightings') produce new and powerful capacities for our parallel-distributed brains (1991a: 201-10). Language use is simultaneously the primary medium for this transmission, and in many cases, the habit of thought itself. In brain terms, meme-effects harness the large amount of adjustable plasticity genetic evolution has endowed, and allows our brains to, in Dennett's words: "take on myriad different micro-habits and thereby take on different macro-habits." The infestation of a successful meme results in "thousands or millions or billions of connection strength settings between neurons" becoming fine-tuned to reflect "a new set of micro-habits, or a new set of conditional regularities of behaviour" (1991a: 218).

Civilization has relied on the practice of developing and transmitting useful habits of thought to others so that we can enhance our communicative repertoires and so that we can benefit from the inventions and hard work of past generations. Since there is no shared machine language between brains that would facilitate the kind of precise data replication a computer is capable of, Dennett characterises human 'software sharing' as social, highly context sensitive, partially self-organizing, and highly format tolerant (1991a: 220). He suggests that if anything like 'software sharing' does occur between human brains that it can be achieved via a number of familiar methods such as "learning by imitation, learning as a result of 'reinforcement' (either deliberately imposed by a teacher – (through) reward, encouragement, disapproval, (or) threat – or subtly and unconsciously transmitted in the course of communicative encounters), and, learning as the result of explicit instruction in a natural language that has already been learned via the first two methods" (1991a: 220).

Rehearsal is the key feature of this process and can be abetted by forming easy to recall maxims and rhymes, by making bodily associations, or drawing pictures and writing texts. In general all these modes of "deliberate repeated juxtaposition of elements" aim at creating useful bridges of association "so that one item would always 'remind' the brain of the next" (1991a: 224-5). The net effect of all this is that each new individual does not have to 'reinvent the wheel' because in Dennett's words: "we *install* an already invented and largely "debugged" system of habits in the partly unstructured brain" (1991a: 193). Dennett argues that human consciousness is "itself a huge complex of memes" (1991a: 210), more specifically, Dennett claims that what makes human consciousness so unique, and so powerful, is not the difference between our brains and those of other higher mammals, but rather the way that the plasticity of our brains is harnessed by language and culture (1991a: 190). It is the effect of these enhancements which is well explained (according to Dennett) as the installation of a von Neumannesque virtual machine in the brain (1991a: 210, 254). In section (1.4) I isolate two important questions that are raised by this proposal, and together, they form the central themes of this essay.

1.2 Numerous Rough Copies

Although I am largely concerned with the second part of Dennett's theory, (the von Neumannesque virtual machine hypothesis), a grasp of the first part, the Multiple Drafts model, is central to understanding the theory as a whole. What follows is a brief characterization of the Multiple Drafts model.

Dennett argues that the brain is a system of 'multiple channels with specialist circuits'. These specialists only operate on their respective tasks, and are not under any sort of global control (1991a: 258). Dennett convincingly destroys the idea that any one part of the brain, or any specialist, is the part that *makes* contents conscious – what he terms the Cartesian Theatre idea. Instead, specialists work in what Dennett calls, 'parallel pandemonium' (1991a: 237-8, 240-2, 251, 253).[2] At any one time specialists around the

[2] This idea originates with Oliver Selfridge (1959).

brain are working on their own 'projects', and 'drafts' (which are produced by the specialists during the fulfilment of their tasks) are 'awarded' consciousness by being comparatively 'famous', or in Dennett's own words "by being temporary winners of the competitions, persisting in the cerebral arena, and hence have more and more staying power (in memory – which is not a separate system or box)" (Dennett 1998: 293). Most drafts only exist for a moment, helping to modulate and control some current behaviour and then fading away. Some, however, get more 'famous' by performing roles that are important or repeated, and these drafts play long-term roles (1991a: 254). The key feature of the model is that content fixations need only be made once – as opposed to being re-represented for the benefit of the audience in the Cartesian Theatre (1991a: 135).

Section (6.2) supplies more details of this model, examines some of its primary functions, and investigates a type of connectionist network that is highly compatible with the Multiple Drafts model.

1.3 Virtual Machines

Recall that Dennett claims human consciousness is "best understood as the operation of a 'von Neumannesque' virtual machine implemented in the parallel architecture of a brain that was not designed for any such activities" (1991a: 210). He describes a virtual machine as "a temporary set of highly structured regularities imposed on the underlying hardware by a program: a structured recipe" (1991a: 216). Lets take a moment to clarify this.

The original notion of a virtual machine is from computer science. A computer can do quite different things when different programs are running on it: at one moment it is a word processor, the next moment a chess machine, an Internet browser, etc. Inside the computer however, you will always find that the same brand of serial processing is responsible for each of these different applications. The word processor, chess machine and the Internet browser are examples of virtual machines running on digital computers. Virtual machines are often simulations of games or equipment that are usually (or

previously) instantiated in devices designed for a single, special purpose. Digital computers installed with a variety of software allow us to have one highly adaptable machine to do the work of many special purpose machines. A chess-playing virtual machine, quite obviously, does not operate over actual chess pieces, but rather over symbolic representations that code for them. The rules for the movement and taking of chess pieces are similarly coded for as algorithms, such that when the virtual chess machine is active, the computer produces an operating environment where the functionally important parameters and rules of transition for chess can be computed, and a user can enjoy a simulation of a chess game.

Connectionist networks are inspired by the way animal brains function, and animal brains function in ways quite different from those of digital computers (see section 1.4 below). However, for practical reasons the vast majority of connectionist research is conducted with digital computers that use software to precisely simulate the functionality of a parallel distributed processing device (or neural network). This is another example of a virtual machine: a virtual parallel-distributed processor. Dennett's theory of consciousness turns this standard idea upside down, and characterizes the collective effects of language use and enculturation as the installation of a serial virtual machine in the parallel processing brain[3] – he states: "Conscious minds are more-or-less serial virtual machines implemented – inefficiently – on the parallel hardware that evolution has provided for us" (1991a: 218).

All examples of serial processors (or digital computers, or von Neumann machines) share a processing technique involving sequential operations over symbolic codes. They all have a similar basic architecture consisting (essentially) of a central processor, a separate, inert memory, and a workspace. Dennett's characterization of conscious minds as von Neumannesque virtual machines (hereafter: vNvm) describes a process whereby human brains, (although parallel processors), come to *simulate* (more or less) serial operations

[3] Importantly, Dennett is not only concerned with an input-output similarity, but with our actual conscious experience and our computational abilities. In other words we make use of symbols, experience a serial, stream of consciousness, have an easily accessed memory, can perform long sequences of operations, apply rules, etc. Section (5.2) will return to the idea of input-output similarity.

on a system of moveable symbols. Importantly Dennett is not claiming that this serial/symbolic process is a description of the brain's bottom line processing dynamics, but rather that the activity of a trained (and still parallel processing) brain comes to simulate, and therefore be describable (from some level of abstraction) as, a von Neumannesque phenomenon. Just as the virtual chess machine allowed for chess-playing without a physical board, chess pieces, and a human opponent, we can (according to Dennett) have a conscious mind resembling a serial, symbol processor without physically existing, discrete symbols, and without denying that the brain is a parallel processor. The next section elaborates on the differences between serial and parallel processing, and sets the tasks for this essay.

1.4 Classicism, Connectionism, and Compromise

Classical cognitive science supports the thesis that human cognition is analogous to the symbolic computation of digital computers. On this account information is represented in strings of symbols just as we represent information in a computer's memory or with written language. Connectionists on the other hand, argue that the brain does not represent information symbolically, but rather, information is stored in the connection strengths between different units in a neural network. Whilst the classicist believes that cognition resembles digital processing (where symbol strings are produced in a sequence according to rules or programs), the connectionist views cognition as patterns of activity over large sets of neurons connected in a network.

In general terms, the symbol-processing models of classical cognitive science seem badly suited for explaining the flexibility and efficiency of human cognition, yet they seem indispensable to an account of tasks like language use and reasoning (Fodor and Pylyshyn 1988). Some supporters of connectionism, like Andy Clark (1989; 1993) and Paul Smolensky (1988), seek an accommodation between the two paradigms by finding ways in which neural networks can implement (some aspects of) symbolic processing. Under this proposal, even though the brain is a neural network, it is also a symbolic

processor at a higher and more abstract level of description. The challenge (which is also part of the focus of this paper: sections 6,7, and 8) is to discover how the machinery needed for symbolic processing can be forged from neural network materials. More radical supporters of connectionism, like Paul Churchland (1989; 1995), claim that symbolic processing was a bad guess about how the mind works, because important features of human intelligence that are well explained by connectionist models (like graceful degradation of function, spontaneous categorization and prototype extraction, and context-sensitivity[4]), are poorly accounted for by the classical approach.

Although Dennett's theory of consciousness is a theory where a connectionist-type brain comes to simulate a symbolic serial process, this paper will not detail the specifics of the many debates between classicists and connectionists, that have been prevalent over the past twenty-five years or so. Dennett's theory entails that aspects of both paradigms are indispensable, and thus, that arguing for the final superiority and ubiquitous applicability of either classicism, or connectionism, is pointless. So instead of investigating the classicist/connectionist debate per se, I want to remain focussed on exploring some substantive issues that concern a theory attempting to integrate the two. As we will see, I am going to accept, with Dennett, that what we need is to understand how a von Neumannesque serial phenomenon can come to exist in the parallel-distributed system of the brain (1991a: 214). But this idea raises a puzzling question: How can a single device be both a parallel and a serial processor? This question leads to two more specific questions that will form the main themes of this paper:

1. Explanatory legitimacy:
   If non-symbolic, parallel-distributed networks accomplish all the representation and computation of the brain, what kind of explanation of the functionality of that brain can legitimately maintain descriptions of procedures that are symbolic, serial, and real?

[4] For a nice survey of the ways in which connectionist networks naturally support these features see Clark (1993), or alternatively, Churchland (1995) or Rumelhart et al (1986).

2. Implementation:

What kind of structural design, training, and resultant processing dynamics could enable a (human) brain to develop a competency for symbolic, serial procedures?

I will be discussing Dennett's vNvm theory in the contexts created by the two questions above. I argue that the theory is generally correct, that Dennett's methodology constitutes a solid answer to question (1) above, and that his theory of consciousness goes a long way towards answering (2). I will address some criticisms and discuss related theories along the way. Sections (2) and (3) will examine the kind of explanation required to deal with the question of explanatory legitimacy. These sections also examine the ways in which Dennett's position on explanation (and content) leads towards the kind of theory of consciousness he has produced. Section (4) looks at a criticism of Dennett's explanatory style, and argues that we should be more concerned with the substance of Dennett's ideas than with the analogical tools he uses to express them. Section (4.3) sets the task for the rest of the paper by isolating four key features of consciousness which Dennett's vNvm theory is designed to explain. Section (5) is slight and necessary detour that uncovers Paul Churchland's explanatory methodology with a view to fielding his many objections to Dennett's theory. In (5.2) I dispense with those objections that appear to be no more than symptoms of the differences between Churchland and Dennett's methods. At this point we will be in a position to discuss the second question cited above. I do so in sections (6), (7) and (8) by examining possible ways in which the four features traced in section (4) could be implemented in the parallel processing brain. During these discussions I will mainly be looking at theories that either compliment or augment Dennett's vNvm theory, whilst highlighting some problems and disagreements. Lets proceed.

Section 2: Explanation

The following sections concern the nature of explanation in cognitive science. My aims here are not to redefine or develop the ideas, and not to supply any great detail concerning relevant and important issues like, for example, the many types, merits, and pitfalls of functionalism, inter-theoretic reductionism, or causality. I aim only to suggest ways in which to address the above question of explanatory legitimacy. Over and above this, these sections will hopefully lead to a better understanding of Dennett's ideas, and establish some useful terminology that I will employ in later sections when we turn to the task of fielding objections to the vNvm idea, and exploring the question of implementation.

2.1 Three Levels

To begin, I will examine a classic analysis of the nature of explanation in cognitive science. David Marr's (1982) tripartite division of the explanatory landscape is designed to apply to any machine carrying out an information-processing task. Here is Marr's condensed description of these levels:

> Level-1 (the computational level): What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
> Level-2 (the algorithmic level): How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
> Level-3 (the implementation level): How can the representation and algorithm be realized physically? (Marr 1982: 25)

Marr uses words like 'what' and 'how' to explain the purpose of each of his levels of explanation. The reason for this is that these different levels amount, more or less, to different *types of inquiry*. The general thrust of Marr's idea is (rightly in my view) that different types of questions demand different types of answers, and comprehensive

explanations involve furnishing answers to all these questions. The relationship between these answers is left fairly open by Marr; he recommends that they be "consistent with one another" or "linked, at least in principle, into a cohesive whole, even if linking the levels in complete detail is impractical" (1982: 20).[5]

Level 1 is the level of explanation for behaviours, functions, competencies, and so forth, in abstraction from physical implementation. Marr stresses the importance of this kind of level (as does Dennett, see sections 2.2 and 3.2) because, as he argues, trying to understand bird flight by studying only feathers is impossible. "In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense" (1982: 27). Further, an algorithm at level 2 is likely to be understood more readily "by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied" (1982: 27). At level 1, Marr tells us, we should have separate arguments for what is computed, and for why this particular computation is appropriate to the goal. In the analogy with bird flight this would involve (1) describing the behaviour of the bird (flying by flapping wings) and (2) an explanation of why 'flapping wings' should be appropriate given the goal 'flying'. This second part then, should make reference to aerodynamics and the resultant appropriate shape of the wings.[6]

Level 1 allows for the isolation of features that adequately define a process, outlining a 'target' for explanations at lower levels. Once this is done we require an explanation of

---

[5] This raises some tough philosophical questions, like, 'are any levels reducible or eliminable'? But Marr's delicate answer above will suffice for my purposes. In other words, I will proceed with the assumption that for complex subjects such as the mind/brain, neat reductions might be impossible either (1) in principle or (2) for all practical purposes. The way this turns out has little bearing on what follows in this paper because as I hope to suggest later (sections 2.2 and 2.3), all Marr's levels (or something like them) remain valuable (and not eliminable) whichever way the reduction question turns out. However, to doubt the possibility if neat reductions is not to say that different 'levels' or 'types of inquiry' are entirely separable from one another, after all they concern the same system or device. For example, as Marr suggests, the answer to the '*what* question', can inform, correct, illuminate, and restrict our answers to the '*how* question'. This is despite the fact that the "explication of each of these levels involves issues that are rather independent of the other two". For example, the choice of algorithm is influenced by what it has to do, and the hardware on which it must run (Marr 1982: 25).

[6] A level 1 account is "defined uniquely by the constraints it has to satisfy" (Marr 1982: 23). The defining constraints in the above example are 'flying with wings' and aerodynamic considerations.

how this process is realized. This brings us to level 2, which involves choosing two things: The first is a representation for the input and output of the process, and the second, is "an *algorithm* by which the transformation may actually be accomplished" (1982: 23). Level 2 (like level 1) remains neutral about the physical materials used and the underlying mechanisms employed – this is the domain of level 3. As we descend through the different levels we lose the width of application that our high-level explanations provided, but gain detailed accounts of how the upper level properties are realized.

Marr's levels are extremely useful but are not without problems.[7] For example, Andy Clark notes that between any two of Marr's levels there are bound to be other theoretically interesting or significant levels. For example, Clark cites Weismannist inheritance as falling between Darwinian theory (level 1) and Mendelian genetics (level 2) (Clark 1993: 43.). The following section takes a closer look at this issue and its relevance to our discussion.

2.2 Using the Levels

An important feature of level-1 theories is that they can take numerous different forms, encompassing different levels of generality, and dictating, to greater and lesser degrees, the contents of lower level theories. As we discovered above, although level 1 explanation counts towards the understanding of a system, it also serves to isolate a process, effectively defining an equivalence class for all things that meet the constraints outlined in the account. If a level 1 account is highly specific it defines a small equivalence class, and vice versa. However, as Andy Clark explains:

> "There is an important gap between the "official" account of the top level (level 1, as Marr calls it) and the actual practice of giving "level 1" theories. Although the official line is that a level-1 account specifies only the what and the why of a

[7] Section (3.2) looks at one or two problems raised by Dennett, and briefly explains the similarities and differences between Marr's methodology and Dennett's.

> computation, this specification can be progressively refined so as to define a more informative (i.e. more restrictive) equivalence class. This more refined version of level-1 theorizing (which yet falls short of a full algorithmic account) has been persuasively depicted by Peacocke (1986) under the title of level 1.5" (Clark 1993: 44).

The relevant contrast here is between "an equivalence class generated by defining a function *in extension* (i.e., by its results – the *what*, in Marr's terms) and a more restrictive (and informative) equivalence class generated by specifying the *body* of information upon which an algorithm draws" (Clark 1993: 44). Clark calls this second type (Peacocke's level-1.5) a competence theory (following Chomsky).[8] Some competence theories are very restrictive, rigidly specifying the kinds of algorithms and representations to be used in level-2 theories. Classical cognitive science is just such an instance. In general, level-1 accounts in such theories are characterized by well-organized symbolic representations, and they use a symbol processing architecture to implement level-2 algorithms, thus, the classicist preserves a close relation between the competence theory and the accounts at lower levels.

> "Indeed, it begins to seem as if that close relation is what *constitutes* a classical approach. Thus, Dennett (1987: 227) visualizes the classicist dream as involving a "triumphant cascade through Marr's three levels." This classicist vision is clearly exemplified in Syntactic Image-style models in which the objects of computation are syntactically structured representations and the computational processes consist of logico-manipulative operations defined to apply to such items in virtue of their structure" (Clark 1993: 46).

---

[8] It is possible to argue that aspects of Dennett's vNvm theory are just such a competence (or level-1.5) theory. To be skeletally brief, if we considered "von Neumanesque" to be Dennett's level 1 account, and emphasised his contention that this competence gets implemented by exposure to language and culture, we could conclude that this constitutes the 'specifying of a body of information from which level-2 algorithms can draw'. However, I don't think this kind of classificatory task will offer us much utility given my broader aims. Later (in section 4.3) I attempt a more useful (for my purposes) decomposition of Dennett's ideas at level 1, followed (in sections 6,7, and 8) by an examination of the feasibility of his suggestions for lower level implementation.

In this context the difference between Dennett's theory and the classicist is that although both describe a level 1 competence for serial, symbolic procedures like language use and reasoning, Dennett claims that "the decomposition of one's competence model into parts, phases, states, steps, or whatever *need* shed no light at all on the decomposition of actual mechanical parts, phases, states, or steps of the system being modelled" (Dennett 1987: 75). This means, for example, that a level 1 description of human consciousness that seemed to accurately isolate a phenomenon with the phrase "two symbols joined together", would not entail that in explaining that phenomenon at lower levels, we search (only) for two originally discrete, physical tokens that have now become fused. In Marr's terms the answer to the 'what question' does not always neatly superpose over the answer to the 'how question', and although this does not mean that Marr's levels cannot influence and restrict one another, it does mean that if we are going to theorize at level 1.5, we had better already know something about the lower level mechanisms in question. The next section supports this point a little further.

Classical theories, by having a close relation between their competence theories and lower level theories have run into trouble because, as Clark puts it: "there is no pervasive text-like inner code", and high-level concepts "do not have inner vehicles in the form of context-free syntactic items in such a code" (1993: 13). The classical competence theory then, demands lower level mechanisms that simply do not occur in real brains. Furthermore, some theorists are finding ways of implementing the level 1 phenomena that classicists rightly deem important by using connectionist methods, (see for example, Elman 1991), demonstrating the claim that, not only are classical-style mechanisms absent in brains, they are also not necessarily required. As I have mentioned, later sections of this work look at some other ways of achieving symbolic, serial competencies within a connectionist setting. For now, I want to return to the dangers of level 1.5, specifically when used in conjunction with a 'top-down' explanatory strategy. If we are to take Marr's advice, and begin our explanations with an isolation of what something does, we should be careful (in refining this description) of what we imply about underlying mechanisms. Lets look at why this is the case.

2.3 How to Change the World from your Armchair

Where I live I don't have a garage in which to park my car. If I did have a garage I should like it to have an electronically operated door. I could then take advantage of a recently developed, and cunning, system of garage door opening (and closing). Normally, electronic garage doors are opened by a small remote control, the new system works as follows: a fixed, remote control-like device is installed into the front of your car (under the bonnet), and is then wired through to the lever for controlling the headlights. This allows you to open and close your garage door by simply pulling on this lever (simultaneously flashing your headlights), and by doing so, eliminates the frustration of losing, or forgetting your remote control. (The headlights, of course, still function as normal.)

Imagine a person who doesn't know anything about electronic garage doors, remote controls, or the innovation described above. Everyday he observes (from his armchair) the following event: his neighbour returns to a garage, flashes his lights, and the garage door then opens. He may then be led (quite understandably) to create the following level 1 theory:

"when my neighbour flashes his headlights the garage door opens"

Perhaps after another month of armchair observations our theorist would attempt to give his theory a bit more detail:

"every time the headlights flash the garage door opens, therefore, *the flashing of the headlights causes the opening of the garage door*"

Our theorist still doesn't know *exactly* how the flashing headlights affect a response from the garage door, but he claims to be able to deduce facts about mechanisms from his behavioural observations, and so moves to turn his theory into a suggestive (level 1.5) competence theory:

"Since the headlights cause the garage to open, there must be some sort of 'light detector' on (or in) the garage."

To deny this would be to deny that the headlights cause the door to open (the central tenet of the developing account), and so our now satisfied theorist, thinks he has solved all the functionally important aspects of the problem. Indeed, we could imagine his convictions to be so strong, that upon unsuccessfully attempting to open this same garage door with *his own car's headlights* he might remark:
"Mmm, the 'light detector' must be faulty!"

This is a case of what Marr calls a failure to recognize the "theoretical distinction between *what* and *how*" (1982: 28). The error above is essentially the same as that which we discussed above pertaining to classical cognitive science, and could be phrased as a problem facing *exclusively* top-down theories. More specifically, this problem faces top-down theories whose upper levels restrict theories at lower levels without sufficiently examining the already existing constraints on these lower levels (hardware constraints). There is a real danger here of such theorists attempting to force the world into compliance with their high-level speculations.

The first garage-door theory made no claims about underlying mechanisms (viz. "when my neighbour flashes his headlights the garage door opens") and was merely a documentation of a common thread evident from many observations. At this point, I believe, we must leave the armchair if we wish to continue theorizing, effectively restricting our level 1 theories to some kind of salient pattern isolation. The crucial mistake made by the imaginary theorist, was to assume that he could discover, from his armchair observations, the (level 2) method used for the implementation of garage door opening.

As we saw above Dennett's denial of any kind of necessary congruency between the elements and events postulated in a competence theory and those of the lower level mechanisms, entails that he is unlikely to make the mistake described above. In *Consciousness Explained* he describes his strategy as follows:

> "...if we succeeded in explaining the process at the level of synapses or bundles of neurons, we would be mystified about other aspects of what must be happening. If we are to make sense of this at all, we must first ascend to a more general and abstract level. Once we understand the process in outline at the higher level, we can think about descending once again to the more mechanical level of the brain" (1991a: 193).

This makes Dennett's strategy a lot like Marr's in the sense that both start by describing what happens, or what it is that needs explaining, but Dennett's contentions at the 'more general and abstract level' are not supposed to be clues about how the mechanical level works – rather just efforts to discover the capacities, functions or behaviours of the system we seek to understand.

We are now in a position to dispose with the most obvious objection to Dennett's vNvm theory, namely: 'the brain processes in parallel so explanations cannot cite systems of moveable symbols'. The response to this sort of objection should now be quite obvious; the idea that different phenomena need to be explained at different levels can allow us to describe 'what a system is capable of' in abstraction from how these capabilities are accomplished. As Marr notes:

> "the distinction between serial and parallel is a distinction at the level of algorithm; it is not fundamental at all – anything programmed in parallel can be rewritten serially (though not necessarily vice-versa). The distinction, therefore, provides no grounds for arguing that the brain operates so differently from a computer, that a computer could not be programmed to perform the same tasks"(1982: 27).[9]

This idea is central to Dennett's proposal, and a similar objection to the one above, has been raised by Paul Churchland (1995; 2002). Section (5.2) discusses this, and

[9] Dennett, of course, *does* think that a suitably assembled and configured parallel-processing device can develop serial competencies.

some further objections from Churchland. The next section briefly outlines Dennett's methodology as he articulates it, explains some of the similarities and differences between Dennett and Marr in this context, and explores (very briefly) the relationship between causality and explanation.

Section 3: Realism and Causality

3.1 Really Mental

I want to begin with a brief discussion of 'mentalistic explanations' because both Dennett and Churchland's positions on these issues are strongly related to their positions on consciousness. Dennett defends a position on such explanations that resists the *perceived* dichotomy between realism and eliminative materialism (1991b: 95). Instead, Dennett argues for what he has called 'mild realism'[10] (1991b: 99; 1994a: 365), a position where the language of folk psychology (and other mentalistic explanations) has a predictive/explanatory power that is objectively real, but resists the notion that beliefs, desires, etc must be (so real as to be) reducible to causally efficacious physical tokens somehow instantiated in brains (see 1987: 13-35). Before supplying more detail here let us quickly examine the extremes of the dichotomy.

Realists (chiefly represented by Jerry Fodor 1975, 1981, 1987, and Fodor and Pylyshyn 1988) argue that mentalistic explanations describe a real causal story, and that as such they must be (at least) token identical to physical, causally efficacious entities. This is nicely captured in Fodor's famous aphorism: "No Intentional Causation without Explicit Representation" (1987: 25). On the other extreme, eliminative materialists (chiefly represented by Paul Churchland 1979, 1981, 1989), argue that folk psychology is a wholly impoverished theory, and that as a result realism is bound to fail. Instead they argue that all mentalistic explanations must be eliminated and replaced by the more precise explanatory language of modern (or future) neuroscience. Churchland puts it

[10] Although a little beyond the scope of this work, it is worth pointing out that Don Ross (2000) has persuasively argued that aspects of Dennett's position here are untenable, as they lead inexorably to reductionism, instrumentalism, or both – (positions with which Dennett would not want to associate.) Ross develops an account of an ontological position ('Rainforest Realism'), which saves Dennett's commitments to the stance theory of intentionality and pattern realism from the uncomfortable consequences above, at the cost of abandoning the distinction (relied upon in Dennett 1991b) between *illata* and *abstracta*. This distinction leads to reductionism or instrumentalism because it invites us to suppose that some entities exist independently of our detecting them, whilst others depend on human constructions. Rainforest Realism is the thesis that we should be pattern realists "all the way down", where patterns are real if "projectable from at least one physically possible perspective"; "more efficient (in information-theoretic terms) than the bit-

succinctly:

> "Eliminative materialism is the thesis that our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be replaced, rather than smoothly reduced, by completed neuroscience" (1981: 206).

My goals here are not to argue for or against realism, mild realism, or eliminative materialism in any kind of detail. Also, I will discuss eliminative materialism and mild realism in much more detail than I will realism. My reason for this is that Paul Churchland is one of the sternest critics of the vNvm idea; understanding both his and Dennett's metaphysical positions here will, I claim, lead to a much better understanding of this debate (in sections 5.2, 6.1, and 6.2).

3.2 Take a Stance

This section provides some details of Dennett's explanatory strategy. Dennett characterizes different kinds of explanations as pitched *from* particular stances, rather than *at* different levels. However, Marr's levels and Dennett's stances have many similarities. The major difference is that Dennett takes seriously the relationship between the explainer and the world she (or it) attempts to explain (Dennett 1987: 39). After I have presented Dennett's ideas, I will discuss further differences between Marr and Dennett's respective frameworks.

Dennett uses the example of a chess-playing computer to clearly illustrate his position, and since it is so clear, I will preserve most of his wording below. If we were attempting to explain or predict the behaviour of a chess playing computer there are three 'stances'

---

map encoding", and where from at least one of the physically possible perspectives there exists some part of the pattern that can only be tracked from that perspective. See Ross (2000) for more details.

we could assume: First there is the *design stance*. If we know how the chess-playing software was designed we can predict the computer's designed response...

> "The essential feature of the design stance is that we make predictions solely from knowledge or assumptions about the system's design, often without making any examination of the innards of the particular object. (...) Second, there is what we may call the *physical stance*. From this stance our predictions are based on the actual state of the particular system, and are worked out by applying whatever knowledge we have of the laws of nature. (...) Attempting to give a physical account or prediction of the chess-playing computer would be a pointless and Herculean labour, but it would work in principle. One could predict the response it would make in a chess game by tracing out the effects of the input energies all the way through the computer until once more type was pressed against paper and a response was printed. (...) There is a third stance one can adopt toward a system, and that is the *intentional stance*. This tends to be the most appropriate when the system one is dealing with is too complex to be dealt with effectively from the other stances. In the case of a chess-playing computer one adopts this stance when one tries to predict its response to one's move by figuring out what a good or reasonable response would be, given the information the computer has about the situation. Here one assumes not just the absence of malfunction but the rationality of the design or programming as well. (...) Whenever one can successfully adopt the intentional stance towards an object, I call that object an *intentional* system. The success of the stance is of course a matter settled pragmatically, without reference to whether the object *really* has beliefs, intentions, and so forth; so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably *are* intentional systems, for they are systems whose behaviour can be predicted, and most efficiently predicted, by adopting the intentional stance towards them" (Dennett 1978: 237-8).

Thus, mentalistic explanations are explanations pitched from an intentional stance, and as

Dennett describes there is no way of establishing a privileged language or area of applicability for such explanations beyond our (pragmatically determined) predictive success rate. For Dennett then, there is nothing *intrinsic* that determines the content of mental states; all content is attributed or ascribed (1991b: 110-20). Humans are especially good at picking out the 'real patterns' in one another's behaviour, in other words we are able to assume the intentional stance. The result of this is that the patterns we observe may not be observable for other creatures (like Martians) but this does not mean that the patterns are any less real (1987: 25-28; 1991b: 110).

Dennett has criticised Marr's approach (particularly level 1) for being too idealistic or optimistic: "what Marr underestimates, however, is the extent to which computational level (or intentional stance) descriptions can also mislead the theorist who forgets how idealized they are" (1987: 28). This point, and the idea that the difference between levels is a difference in the approach or perspective of the explainer, (rather than a more objective difference), are the two main points of departure between Dennett and Marr (on explanation). Similarly, Dennett sees the difference as one of "emphasis" (1994b: 253), and notes the three main similarities:

1. stress on being able (in principle) to specify the function computed (...) independently of the other levels;
2. an optimistic assumption of a specific sort of functionalism: one that presupposes that the concept of the function of a particular cognitive system or subsystem can be specified (it is the function which is to be optimally implemented);
3. A willingness to view psychology or cognitive science as reverse engineering in a rather straightforward way (1994b: 253-4).[11]

I will not be attempting an evaluation of these three points in this work. I hope nevertheless, to have provided a characterization of Dennett's method that allows us to proceed with an understanding of the metaphysical commitments lurking behind the

[11] Reverse engineering is the interpretation of an artefact (or even an organism) "by an analysis of the design considerations that must have governed its creation" (Dennett 1994b: 254).

vNvm hypothesis. *Consciousness Explained* does not supply much *explicit* detail on Dennett's position with respect to levels of explanation (nor his position on content), and this is quite understandable given that the aims of the project were to develop a theory of consciousness that could dissolve some bad intuitions, and sort out some messy philosophical debates. However, Dennett's position in the book is more readily understandable, and less likely to appear strange, if it is read within the context of his broader commitments. (For example, some people (myself amongst them), upon encountering the book for the first time without much of an understanding of Dennett's philosophy, may have expected a book with such a title to contain rather more neuroscientific evidence.) I hope that the preceding sections will have such readers thinking about which level/stance Dennett has assumed to explain consciousness. I supply my own interpretation shortly, in sections (4.1) and (4.3). Before this however, one more aspect of explanation cannot escape our attention. Up to this point I have avoided any sort of detailed discussion of causality and its relation to explanation. The next section supplies a minimal characterization of the way in which methodological frameworks like Dennett's and Marr's incorporate causal explanations. I will do so, not by discussing either author's work directly, but through a discussion of Andy Clark's perspective on this issue, which I take to be highly compatible with both frameworks.

3.3 Causal Efficacy

According to Clark the condition of causal efficacy states: "a psychological ascription is only warranted if the items it posits have direct analogues in the production (or possible production) of behaviour" (1989: 196). He goes on to provide a neat summary of the eliminativist argument:

> *Step 1.* Suppose that pure distributed connectionism offers a correct account of cognition.
> *Step 2.* It follows that there will be no discrete, recurrent, in-the-head analogues to the conceptual-level terms that figure in folk psychological belief ascription.

> *Step 3.* Hence by the condition of casual efficacy, such ascriptions are not warranted, since they have no in-the-head counterpart in the causal chains leading to action.
> *Step 4.* Hence, the casual explanations given in ordinary terms of beliefs and desires (e.g., "She went out because she believed it was snowing") are technically mistaken (1989: 196).

Clark argues that even if pure distributed connectionism offers the correct account of cognition, the eliminativist conclusion (step 4) does not follow. The reason for this is that "good causal explanation in psychology is *not* subject to the condition of causal efficacy" (1989: 197). (The reasons for this are discussed below.) Clark then extends this move to symbolic AI, viz. even if pure distributed connectionism offers the correct account of cognition, it does not follow that symbolic AI is false, approximate, or useless. Instead, he argues, symbolic AI (and Folk Psychology) "are geared accurately to capture legitimate and psychologically interesting equivalence classes, which would be *invisible* if we restricted ourselves to subsymbolic levels of description" (1989: 197).

Interestingly, (like eliminative materialists) realists, such as Fodor and Pylyshyn also make the condition of causal efficacy pivotal in their defence of folk psychology and symbolic AI. Their perspective on levels of explanation can be inferred from the following:

> "It seems certain that the world has causal structure at very many different levels of analysis, with individuals recognized at the lowest levels being, in general, very small and individuals recognized the highest levels being, in general, very large. Thus there is a scientific story to be told about quarks; and a scientific story to be told about atoms; and a scientific story to be told about molecules ... ditto rocks and stones and rivers ... ditto galaxies. And the story that scientists tell about the causal structure that the world has at any one of these levels may be quite different from the story that they tell about its causal structure at the next level up or down. The methodological implication for psychology is this: if

> you want to have an argument about *cognitive* architecture, you have to specify the level of analysis that's supposed to be at issue." (Fodor and Pylyshyn 1988: 95)

In the pages following this quote Fodor and Pylyshyn name the *cognitive level* as the one appropriate for psychology, and this level is supposed to be continuous with the scale suggested (from quarks and smaller, to galaxies and beyond). Since this scale consists only of causally efficacious entities, Fodor and Pylyshyn attempt to argue for neat, 'in the head' correlates to symbolic descriptions. Without doing so they would be forced to give up one of the following claims: (1) all explanations must describe causally efficacious entities only, or (2) folk psychology (and symbolic AI) track causally efficacious entities. Notice that the eliminativist abandons the second claim, but agrees (with the realists) about the first claim.

Dennett and Clark disagree with both claims. Instead as we saw, Clark argues for "a notion of causal *explanation* without causal *efficacy*" (1989: 197). In other words not all explanations must (have the potential to) be precisely mapped or reduced to fundamental physical causes, in order to be legitimate, useful, and illuminating. Rather, we can create valid models of psychological phenomena by making generalizations over large sets of physical causes. Furthermore, without making such generalizations we will be unable to track really existing, illuminating *patterns* that are crucial to a complete understanding of a system or device. Lets look into this.

Clark notes that Frank Jackson and Phillip Pettit (1988) develop just such a notion and their examples illustrate the case nicely[12]: "Electrons A and B are acted on by independent forces Fa and Fb respectively, and electron A then accelerates at the same rate as electron B. The explanation of this fact is that the magnitude of the two forces is the same. (...) But this sameness in magnitude is quite invisible to A (...) This sameness does not make A move off more or less briskly" (1988: 392-393). And also: "We may

[12] Dennett has put forward many examples that make similar or related points to the ones I discuss above, see for example, the discussions of centres of gravity, 'frames', and Conway's Life, in Dennett 1991b.

explain the conductor's annoyance at a concert by the fact that someone coughed. What will have actually caused the conductor's annoyance will be the coughing of some particular person, Fred, say" (1988: 394).

Lets discuss the latter example first. Suppose, as Clark does, that somebody points out that a more accurate explanation, even, the "proper (fully causal) explanation of the conductor's annoyance was in fact *Fred's* coughing. There is a good sense in which their 'more accurate' explanation would in fact be *less powerful*" (1989: 197). The reason for this is that the explanation that uses "someone" makes it clear that the coughing of *any* audience member would have caused the annoyance in the conductor. Similarly in the electron case, the "sameness", which plays no causal role in the acceleration of either electron, highlights the fact that an infinite number of values of Fa and Fb would result in the same acceleration in A and B, provided Fa is identical with Fb.

Clark describes this increase in generality as an 'explanatory virtue', for although it comes at the cost of "sacrificing the citation of the actual entity implicated in the particular causal chain in question" (1989: 197), it allows for the generation of useful, irreducible, and illuminating equivalence classes which are absent from explanations that are restricted to causally efficacious processes only (1989: 199-201). Clark then, identifies two distinct kinds of explanation:

1. "explanation that highlights common features of a range of cases but abstracts away from the causally active features of a particular case"
2. explanation that "cites the very features that are efficacious in a particular case or range of cases" (1989: 198)

Jackson and Pettit (1988) name the first kind a *program explanation* and the second a *process explanation*. With program explanations "the common feature or property will be

said to causally program the result without actually figuring in the causal chain leading to an individual action or instance" (Clark 1989: 198).[13]

Rather than arguing fully for program explanations or just assuming their metaphysical legitimacy, I hope to have merely suggested the plausibility and, perhaps necessity, of this kind of explanation. In Marr and Dennett's explanatory systems, process explanations only occur at level 3 and the physical stance respectively. I want to resist making the all too simple claim that all other levels/stances are thus program explanations, but it does seem that they are 'causal explanations without causal efficacy' which benefit from the virtues Clark points out.

More importantly Clark's strategy with regard to symbolic AI (above) is almost indistinguishable from Dennett's strategy in Consciousness Explained, where his term 'von Neumannesque' appears to capture a legitimate and illuminating equivalence class, although it does not describe a causally efficacious symbolic, serial processor. In the sections that follow I start to make use of the methodological concepts we have established thus far, to examine more specific aspects of Dennett's theory of consciousness.

[13] Ross and Spurrett (forthcoming) argue that program explanations can be regarded as causally efficacious. In this they disagree with Kim, and Jackson and Pettit themselves. Their argument depends, among other things, on distinguishing scientific from metaphysical problems regarding causation, and pointing out that the view that program explanations lack efficacy depends on confusing the two.

Section 4: Style and Substance

4.1 Dennett's theory within the framework

This section divides the core aspects of Dennett's vNvm hypothesis across the explanatory framework I have discussed in order to set the remaining tasks for this project. In terms of the preceding sections we should expect an explanation of human consciousness to answer the following questions, using the (indicated) appropriate level or stance to do so:

1. What are the features of human consciousness that are most salient and fundamental? (a "what does it do" question) – Marr level 1/intentional stance[14].
2. What methods (programs/algorithms/training) could conceivably support these competencies? (a "how is it done" question) – Marr level 2/design stance.
3. What 'hardware' do we have available, and how does it work such that it implements the employed methods/algorithms? (a question of materials and physical mechanisms) – Marr level 3/physical stance.

The third question is largely the domain of neuroscience and neither Dennett's theory (nor this paper) provides much detail at this level. (It does seem though, that the super-brief answer here is: parallel distributed processing using the billions of neuron cells, in interaction with chemical neurotransmitters, in human brains.) The answer to the first question is partly captured by Dennett's term 'von Neumannesque'; section (4.3) below is entirely devoted to this question, so I will not discuss it here.

---

[14] One examiner points out that my construal here implies that this level is restricted to offering *descriptions*, whereas for Dennett, *explanations* can be pitched from this level, (more specifically, explanations that track the ways in which cognitive and behavioural tasks are shaped and constrained by biological and cultural evolution.) As a future note will mention, the use of my original construal partly weakens the argument of section 5.1, however, the next section (especially: 36) does ask for real ontological commitment at level 1.

A large portion of Dennett's theorizing occurs at the design stance (level 2). As we saw in section (1.3) the von Neumannesque features of consciousness are simulated by 'software', but this tells us more about the development of this simulation than how it is actually supposed to work. We are given clues: Dennett suggests a 'pandemonium-style' implementation with his Multiple Drafts model, but further details are harder to pin down. Sections (6) through (8) are devoted to an attempt to understand (from the design stance) how, according to Dennett, the parallel-distributed brain comes to simulate a von Neumannesque virtual machine (supplying some complimentary suggestions from other theorists along the way).

Before the above-described exploration begins, I look into a possible criticism of Dennett's explanatory style in Consciousness Explained. 'Style' I take to be distinguishable from the explanatory frameworks I discussed in sections (2) and (3), and I mean the word to refer to the particular ways in which a theorist communicates his ideas. This should become clear in what follows.

4.2 Is Style Important?

Dennett's von Neumannesque virtual machine hypothesis is grounded in the idea that the functional properties of a serial/symbol processor can be realized without *actual* symbols, and literal, linear sequences of basic operations upon them, hence Dennett's use of what I call the 'virtual relation' in his theory. The idea that consciousness is a virtual machine *of any kind* serves two purposes (both of which we have touched on above):

1. The first is a metaphysical point concerning Dennett's contention that "the decomposition of one's competence model into parts, phases, states, steps, or whatever *need* shed no light at all on the decomposition of actual mechanical parts, phases, states, or steps of the system being modelled" (Dennett 1987: 75). (The analogy with computer science holds here because, for example, claiming that a computer has a virtual machine that 'plays movies' does not mean we should expect to find film, or a projector, inside the computer's machinery.)

2. The second purpose is to capture Dennett's claim that the competencies described by the term 'von Neumannesque' are ones humans *acquire* (primarily through linguistic activity) rather than being innate. (The analogy again holds in this case: The useful applications of a computer's virtual machines are not available through being 'hard-wired' into the hardware, but rather require the installation of software.)

The criticism I want to consider here (and then largely rebut) is that the analogy of a 'virtual machine' is unnecessary and misleading once the two claims above have been established. Without the 'virtual' construal (it could be argued) Dennett's theory would be better suited to play a guiding role for lower level explanations. The argument could run along the following lines:

We are required to explain a simple hi-fi system that plays a cassette recording of a jazz band. We might begin by describing the behaviour of the hi-fi as 'playing jazz music'. We could then provide analogues of the two claims above, and explain (1) that the usual ways of producing jazz – a literal jazz band – are not present inside the hi-fi, and (2) that the manufacturing process of a hi-fi is not sufficient for the production of jazz music, but rather, a cassette which has been subjected to a recording process can be used to make the hi-fi play jazz (the software analogy). To then claim that the hi-fi *instantiates a virtual jazz band* is simply superfluous – it adds nothing that has not been said more clearly already, and if anything, invites confusion by condensing our important claims. If we were trying to explain the lower level mechanisms of the hi-fi system, we are better served by the two clear claims above, than by a single claim which only *implies* the first two. Furthermore, someone seeking lower level explanations may think the claim that human consciousness is a vNvm only implies *one* of the claims above, and thereby make an error resulting from ignorance of the other. The argument could continue: Another way of illustrating what makes the virtual relation confusing is to consider how different the two claims really are: one describes Dennett's metaphysical views on the relationship between a competence theory and lower levels of explanation, and the other is a claim about the development of human consciousness. It is not necessarily the case that

Dennett's claims about the nature of explanation, and the development of a serial, symbolic consciousness, hang and fall together, so what justifies the move to pull the two together under one analogy? Finally, a further implication of this 'blended' explanation is that it is easy to construe the von Neumannesque competencies as 'virtual competencies', whereas what Dennett seems to mean is that the very real competencies are *supported by* a virtual machine. The virtual machine I am currently using *literally is* a word-processor. Humans can *literally* perform serial-like operations like long division and deduction. The music emanating from the hi-fi literally is jazz music. Competencies documented at level 1 are not virtual competencies.

I do not think any of this criticism would worry Dennett much. I find it unlikely that someone wishing to discover the lower level mechanisms supporting human consciousness would make use of *Consciousness Explained* by focussing on Dennett's analogy *only*. Furthermore, they could extract the two clear claims above quite easily (as I have done); Dennett is not vague about either of them when he unpacks his analogy. Furthermore, Dennett does not need any more justification for mixing the two claims beyond explanatory efficiency, and if one of the claims turned out false, the analogy would be poorer but the remaining claim would not be affected, and could thus be reiterated in some other way. Lastly, it is true that the competencies Dennett deems important are real, but the two claims for which the hypothesis is created do not imply the contrary, to think otherwise would merely be an exegetical mistake, and not something to hold Dennett accountable for. Despite all the above, I do find that one (albeit less important) criticism does hold. Thinking of consciousness as a virtual machine is superfluous (or perhaps: 'dispensable') if you already understand and accept the two claims. You could use whatever schema you felt was superior so long as it was faithful to the core commitments. In other words if you believe the two claims above are true, it does not follow that the *only* way to understand consciousness is as a virtual machine. Indeed, Dennett makes no such claims; he just asserts that it is the *best* way to understand consciousness (1991a: 210).

As we shall see in detail later Paul Churchland disagrees with aspects of Dennett's theory of consciousness, and it appears that he (additionally) does not enjoy Dennett's explanatory style (1999: 763; 2002: 68; 78). What I hope to have suggested with the foregoing section is that those cognitive scientists who share Churchland's stylistic complaints (see section 5.2 for more on this) should not be too hasty in throwing out the bathwater. More specifically, Dennett's *way* of explaining consciousness is separable from the important claims he makes in those explanations. Readers of Dennett who are troubled by his style should, in my opinion, focus on the essential and substantive claims (like the two above) rather than worrying about the packaging.

It is now time to turn towards an examination of Dennett's substantive commitments. The next section begins this task by attempting a level 1 account, not of human consciousness as a whole, but of the phenomena Dennett aims to account for with his vNvm theory. The result is four features that together constitute my interpretation of what Dennett means by his term 'von Neumannesque'. These four features then form the framework for (most of) the remainder of the paper, where I will be addressing the second question I isolated in section (1.4). The four features I isolate are those phenomena, which I argue, we must account for in parallel-processing terms, if we are to successfully explain how a von Neumannesque phenomenon can come to exist in the brain.

4.3 The Competence of Consciousness

Essentially Marr level 1 aims at "understanding the nature of the problem being solved" (Marr 1982: 27). But, Marr level 1 descriptions can differ from theorist to theorist in the same way as two novelists can use very different descriptions of the same type of event. Nevertheless, it is hoped that over time we can come to agree (more or less) on a basic and neutral description of the most fundamental phenomena in need of explanation. For human consciousness this task is rather difficult because of the following two possibilities: (1) it could be the case that humans across different cultures, age groups, (and perhaps even genetic traits, genders, food sources, etc) could have conscious lives of various *types*. And (2), when involved in different activities an individual could shift

between different types of consciousness. These are possibilities worth keeping in mind, and so perhaps the following characterization should be said to suit *most* instances of human consciousness, *most* of the time. And even if this is too strong a claim, the phenomena I list below are at least present in some humans, and more importantly given the task we are engaged in, they are present within *Dennett's* characterization of human consciousness.

The brain it seems can support (what we experience as) related sequences of representations (although not always *obviously* related). Sometimes we can be more specific about the character of such a series of representations, for example we may describe one such instance as a narrative concerning some past event; or a wish for the future; or a real-time description of our forming perceptions; or a chain of seemingly arbitrary associations; or a constantly unfolding creation; or an ordered and deliberate computation; or some speedy, routine steps to some currently needed information; or a synthesis of presently perceived data and our own internal commentaries; or a rendition of an aphorism; or a recital of a list; or a purposeful search for a distant memory; or even, just one part of a familiar song, over and over again.

This list only scratches the surface. In addition to many variations of the above, and the many hundreds of descriptions missing from it, conscious representations can take many forms. For example, our fond recollection of a day in the past, could appear as a string of words, or pictures, or colours, or sounds, or smells, or tastes, or symbols, or tactile sensations, or emotions, or any combination of these and other unmentioned categories. There are however at least three common features to all the conscious activities on our list: (1) They are representational; (2) these representations occur in a non-discrete sequence or stream; and (3) there seems to be some order to these sequences such that they accurately represent the world, or *mean* something to us, or are useful for some purpose, or represent what we require, perceive, think, etc, as opposed to being arbitrary, incongruous, or incomprehensible. (Even though this *is* the case sometimes.) I consider these three features to be relatively uncontroversial, and thus will not waste time arguing for them, indeed as George Miller suggests, my claims are relatively widely accepted:

> "No doubt everyone agrees that consciousness is filled with well-organized objects, events and symbols; consciousness refers to things that exist outside itself. Everyone agrees that the content of consciousness changes from one moment to the next. And we must all agree that we cannot think about everything at once. Everyone would probably agree that changes in consciousness are continuous, there is no succession of discrete photographs, but a stream that flows from one state of consciousness to the next" (1962: 54).

Dennett's theory of consciousness attempts to explain these varied features. However, in order to make our explanatory target more specific and manageable, I want to divide it over three central, and interrelated trends that emerge both from the preceding sketch and from Dennett's 'von Neumannesque' characterization. Here then is (for my purposes) a Marr level 1/intentional stance account of four central problems which Dennett's theory of consciousness (and more particularly the vNvm idea) is designed to explain. There may indeed be more aspects to the theory, and most certainly, there is more to human consciousness, but this list is just a characterization of the problem space for the rest of this paper:

1. Sequentiality: this is my term for the serial, 'stream-like' nature of consciousness. As section (6.2) shows, I use this term because it is neutral about other aspects of consciousness (for example whether consciousness is ordered/controlled or symbolic). Sequentiality refers to the way in which human consciousness flows from one thing to the next, or as Dennett puts it, "your only access to what is going on in your brain comes in a sequential 'format'..." (Dennett 1991a: 215). Sequentiality does not refer to the nature of the contents of consciousness at all beyond the 'successive' character in which contents are conscious. It is thus about how consciousness continues, rather than about what the contents of consciousness can represent, accomplish, etc. Section (6) examines Churchland and Dennett's proposed explanations for this phenomenon.

2. Awareness: Explaining some sort of system that could support a sequential phenomenon is not enough; we must explain the capacity for an 'awareness' of this stream (Dennett 1991a: 189, 225-6, 281). Sections (6.3 and 7.6) deal briefly with this idea, and its relation to the discussion of our third feature.
3. Symbols: Consciousness in adult humans seems to have a strong linguistic influence, such that (some of) the contents of consciousness come to resemble a system of 'moveable' symbols. We have the (conscious) ability to isolate certain contents, so that they can be treated almost as if they were transportable 'objects'. We can extract information from the context in which it was learnt, recombine this with other elements, and produce (a theoretically open-ended) set of novel contents. This feature is the domain of sections (7.1-7.4).
4. Control: Consciousness is (sometimes) characterized by ordered, goal-directed, controlled, or organized episodes (Dennett 1991a: 228, 277-80). In other words, 'order' refers to the capacity to think the right thoughts at the right times (1991a: 222-3). The final section of this paper (8) briefly explores this topic.

It is perhaps with all four of these inter-related features that we are able to be such adept language users, and perform episodes of 'serial reasoning'. When performing serial reasoning (or a serial search) humans are perhaps being their *most* von Neumannesque. It involves features strongly related to serial machine processing: checking or manipulating a sequence of data structures according to a rule (or set of rules). By using this process a large and complex problem can be tackled (very effectively) by decomposing the task into a sequence of incremental steps (involving smaller and simpler problems). Completing the sequence of more tractable tasks eventually yields the solution to the original complex problem.[15] But as Dennett writes, "the standard trap is to suppose that the relatively rare cases of conscious practical reasoning are a good model for the rest, the cases in which our intentional actions emerge from processes into which we have no access" (1991a: 252). I don't claim to have described an all-inclusive competence theory above, and so provided we keep Dennett's warning in mind we can see our way round the

[15] In section (8.2) I will discuss some ideas which suggest that our capacity to perform this kind of reasoning begins by using our surrounding environment to make the task easier.

trap. The explanations of the features above are strongly related in Dennett's account. It is possible however, to discuss each one (fairly) separately without repeating oneself, and this is what I intend to do. I do so in just the order they are written in the list above. However, before we can tackle the first problem, we require a slight detour.

Within the discussion of our first feature (sequentiality) I will also be addressing a number of criticisms that Churchland has levelled against Dennett's ideas, and examining Churchland's own theory of consciousness (with respect to sequentiality). For this reason we need to understand Churchland's method from the outset, and also set aside those criticisms which seem to be symptoms of the differences between Dennett's method and his own. Once this is done we will be equipped to begin our exploration of some ideas for how the four features above might be implemented in human brains.

Section 5: A Virtual Disagreement

5.1 Churchland's Method

In terms of the explanatory perspective described in the earlier sections, folk psychological descriptions are plausibly (albeit, quite awkwardly) interpretable as Marr level 1-type explanations: They do not tell us anything about underlying mechanisms (at least for a mild realist or eliminative materialist) but merely present a description of *what* happens and *why*[16] (at least for the realist[17] and the mild realist). For instance, if *what* happens is the *action,* 'Joe goes to the shop', then *why* this happens could be 'Joe *desires* milk; Joe *believes* milk can be purchased at the shop'. For the eliminative materialist of course this is an impoverished account, one that does not draw sufficiently on modern neuroscience, but how far do the claims of the eliminative materialist extend? More specifically, within the framework I have discussed, do eliminative materialists expect completed neuroscience will serve to eliminate (1) folk psychology only, or (2) Marr level 1 in entirety? I will examine this question only as a bridge to our more important agenda: My main goal in this section is to uncover Paul Churchland's explanatory methodology.

I think Churchland is quite right to point out that folk psychology is sensitive to a small part of mental phenomena:

> "As examples of central and important mental phenomena that remain largely or wholly mysterious within the framework of folk psychology, consider the nature and dynamics of mental illness, the faculty of creative imagination, or the ground of intelligence differences between individuals. Consider our utter ignorance of the nature and psychological functions of sleep, that curious state in which a third of ones life is spent. Reflect on the common ability to catch an outfield ball on the run, or hit a moving car with a snowball. Consider the

[16] Incidentally, this holds nice parallels with Andy Clark's view of folk psychology as providing "a vocabulary with which to specify some of the *targets* of the cognitive scientific endeavour" (1993: 4).

> internal construction of a 3-D visual image from subtle differences in the 2-D array of stimulations in our respective retinas. Consider the rich variety of perceptual illusions, visual or otherwise. Or consider the miracle of memory, with its lightning capacity for relevant retrieval. On these and many other mental phenomena, folk psychology sheds negligible light" (Churchland 1981:210).

Perhaps the main reason for this is that folk psychology has not been carefully developed to play an explanatory role in cognitive science, but rather (perhaps) to help humans understand and predict one another's behaviour. These speculations are inconsequential to my argument here; I wish only to draw attention to the other competencies that Churchland describes, and this, because they appear to be partial descriptions at Marr level 1. From the context of this quote it would seem that Churchland has no problem with deploying (at least partial) high-level descriptions of phenomena, in principle, but only with those particular level 1 descriptions postulated by folk psychology.

Another example (from a different context) gives this view more support: When *engaging* in explanations (rather than philosophising about them) Churchland again uses Marr level 1-type accounts; this time to guide his neuro-scientific speculations. For example, his theory of consciousness begins by setting out what he calls, "the salient dimensions of human consciousness":

1. consciousness involves short term memory
2. consciousness is independent of sensory inputs
3. consciousness displays steerable attention
4. consciousness has the capacity for alternative interpretations of complex or ambiguous data
5. consciousness disappears in deep sleep
6. consciousness reappears in dreaming, at least in muted or disjointed form
7. consciousness harbours the contents of the several basic sensory modalities within a single unified experience (1995: 213)

---

[17] If you were a realist, of course, folk psychological 'derivatives' would feature at Marr level 2 too.

Each of these is followed by a brief characterization of the feature in question, for example, item number 4 reads: "Once one's attention is fixed, on a particular visual scene, for example, a conscious person is still able to generate and explore competing interpretations of the contents or the nature of that scene, especially if the scene is in some way confusing or problematic" (Churchland 1995: 214). Churchland cites these features in an effort to clarify the problem space that theories of consciousness present, providing in his words: "a provisional explanatory target" (1995: 213) for a scientific neurocomputational theory of consciousness. Thus we may speculatively conclude that, if the main function of Marr level 1 is to "understand the nature of the problem being solved" (Marr 1982: 27), then Churchland most certainly makes use of this level (in practice).

It is possible that Churchland objects only to Marr level 1 accounts that are framed as 'explanations' rather than mere 'targets' for the *real* (neuroscientific) explanations. Or perhaps he objects just to those Marr level 1 accounts which claim to be tracking a causally efficacious process by virtue of their predictive power, (or more generally, just those which make demands on lower levels of explanation). Either way, it seems possible to me that if asked to accept the construal (of Marr level 1) over the foregoing sections (as an explanatory target, or problem space isolator), that Churchland would not put up much resistance.

However, this would put some strain on Churchland's eliminative materialism, because if, as he seems to demonstrate through his method, we do still need Marr level 1 guiding explanations, (and folk psychology can usefully be seen as playing this role), then he is not eliminating folk psychology, but rather replacing it.[18] Furthermore, if Churchland accepted the level 1 construal I have outlined he is left with the burden of demonstrating why folk psychology remains so badly suited to play (even a partial) targeting role. In

[18] Interestingly, Churchland has considered this possibility in his *Matter and Consciousness*, where he describes a less extreme form of his position called "revisionary materialism" which advocates, (rather than wholesale elimination), a process whereby only some folk concepts would be eliminated, and others would be improved, replaced and reduced (1988: 49).

other words, what is the (explanatory) difference between claiming that we need to explain 'steerable attention', and claiming we need to explain 'desiring milk'?[19]

Unfortunately in light of more pressing matters I have to leave this issue somewhat unresolved, nevertheless, the foregoing discussion has revealed the important (for my purposes) aspects of Churchland's methodology. We could summarize it as involving two steps:

> Step 1 – A brief, high-level, description of a phenomenon that serves to isolate and clarify the problem space for Step 2.
> Step 2 – The phenomenon (captured by Step 1) is explained by reference to the causally efficacious neural machinery responsible for its existence.

Importantly, Churchland leaves no space for a program explanation (or Marr level 2, or the design stance) primarily because, as we briefly discussed (in section 3.1 and3.3), causally efficacious explanations are the *only* explanations for Churchland, and this level fails his stem test. Thus, in terms of his theory of consciousness, Churchland proceeds by moving from level 1 directly to level 3. Dennett and Densmore criticise Churchland for the absence of such an "intermediate" level, claiming that without this he is powerless to address certain central phenomena, which are visible only as "higher-level patterns of activity" (1999: 750). Churchland on the other hand criticises Dennett for not giving sufficient theoretical attention to some promising aspects of neuroscience and connectionism (1995: 264-9).

If we accept that explanation should proceed at all three of Marr's levels (or something like them), then the above arrangement could seem like a nice division of labour. This is especially the case in the context of consciousness because (as we shall see) there is a large degree of agreement between Dennett and Churchland about what the problem

[19] An examiner pointed out that this argument suffers from the way I characterized level 1 earlier (see: 33). The reason for this is that if there is no ontological commitment at this level then, of course, Churchland can't (or perhaps, would have no reason to) reject it. The examiner recommends that the important issue for

space involves here (level 1). However, a puzzling debate has developed between the two, over the usefulness of the vNvm concept. In super-short, the problem is as follows: Dennett recognises Churchland's primary level of explanation (level 3) but seems to think that it is unimportant as all the interesting phenomena are only visible at level 1 *and* 2, whereas Churchland does not recognise the value of Dennett's explanations at all. The following sections look at the details of the debate.

5.2 Unproductive Objections

Churchland's criticism of vNvms runs something like the following: He thinks Dennett is covertly committed to an outdated language based model of the mind (what he calls 'the classical prototype'), and that he does not give due attention, or credit, to our newfound and powerful knowledge of dynamically recurrent neural networks (1995: 266). In his own words Dennett is trying to "pull a classically serial rabbit out of a massively–parallel human hat" (1995: 267). Later (section 6.1), I will briefly discuss part of the potential of Churchland's recurrent networks to explain consciousness, but for now I want to concentrate, not on his proposals, but rather on his criticisms. I consider the following to be symptoms of the differences between Churchland and Dennett's metaphysical commitments.

Churchland asks: How should we best describe a simulation of a PDP network by a von Neumann machine? And in answer claims: 'all we get from the serial machine is the abstract input-output behaviour of the network being simulated' (1995: 266). (The serial machine is never involved in any actual distributed coding or parallel processing.) And so similarly, Churchland claims, in a simulation of a von Neumann machine by a PDP network, nothing more is simulated than the abstract input-output behaviour of the von Neumann machine (1995: 267). The PDP network is not involved in any actual rule governed, discrete state processing. Of course, from a mechanistic perspective, this is certainly true and by using the phrase 'all you get *from the serial machine*' Churchland

---

Churchland (and other eliminativists) is whether or not level 1/intentional stance explanations lead us to ask the right questions about evolutionary and social dynamics.

makes sure this is the level of analysis we imagine. However, what we get from observing the workings of the serial machine is not *all* we get. Also available (to a user of PDP simulation software) is an interface where network architectures can be designed, training regimes set up, connection strengths observed, etc. In short there is also a program level, and at this level, much more is available than mere input-output similarity. Although there is no PDP in the casual story leading to the outputs produced, users really can adjust weightings between nodes, reorganize the training set, switch off back-propagation, etc.

As Churchland rightly points out, Dennett's account requires more than an input-output similarity, but the following quote suggests a rather confused perspective on this: "The parallel system of the brain must realize genuinely serial computational activities, else the required Joycean stream of consciousness will be strictly absent. Dennett's discussion of "virtual machines" and "simulation" is therefore not to the point. To get what his core theory requires, Dennett needs to find *real* serial procedures within the biological brain" (1995: 267) This comment is false, and quite peculiar given the patent assertions in conflict with it throughout Consciousness Explained (Dennett 1991a: 210-1, 216-21, 225-6, 258-9). A lot here rests on Churchland's phrase '*real* serial procedures'. Dennett would be quite happy to admit that the serial procedures he is speaking of (in human consciousness) are perfectly real - real patterns observable from a high level of abstraction. But Churchland has in mind a stronger sense of the word, which implies that Dennett needs the brain to somehow *change into* a literal, physical, serial processor. However, Dennett does not need to find literal serial procedures, because he is only postulating virtual serial procedures - all Dennett needs, is to find ways in which the parallel-processing brain can support serial, sometimes symbolic, and sometimes controlled (conscious) competencies. Since Churchland does not recognise the 'virtual machine level' of explanation, he asserts that without a literal, serial processing brain the required seriality will be absent, but perhaps he hasn't considered the implications of this comment. If he stands by this claim consistently he must also agree that without real PDP processing (real neural networks), the patterns observed (by PDP researchers) in the digital computer simulations of neural networks are similarly absent, because all the

relevant indications of parallel processing are only visible at this virtual machine level. Yet we don't find Churchland claiming that the people conducting PDP simulations need to find 'real *parallel* procedures' in their digital computers lest their parallel-processing research be 'strictly absent'. As the preceding assertion is directly analogous to the one levelled against Dennett, it would appear (that in this context) Churchland is guilty of a double standard.

In light of his assertion that Dennett needs "real serial procedures in the biological brain", Churchland describes the following as an irony:

> "The fact is, if we do look to recurrent neural networks – which brains most assuredly are – in order to purchase something like the functional properties of a von Neumann machine, we no longer *need* to 'download' any epigenetically supplied meme or program, because the sheer hardware configuration of a recurrent network already delivers the desired capacity for recognizing, manipulating, and generating serial structures in time, right out the box. (...) the need for a virtual vN machine (...) has now been lifted. The brain doesn't need to import the 'software' Dennett contrives for it: Its existing 'hardware' is already equal to the cognitive tasks that he (rightly) deems important." (2002: 71-2)

Importantly, this suggests some consensus at level 1,[20] but more to the point, this quote highlights the tension between Dennett and Churchland's methodologies, which I outlined above. Churchland seems to think that his theory is in direct competition with Dennett's, but his speculations are not pitched at the same level of explanation. Thus, it is conceivable that the mechanisms underlying Dennett's program, *could be* precisely the recurrent PDP networks Churchland cites (more on this towards the end of section 6.2). The debate then, seems to look as pointless the following:

---

[20] Dennett, incidentally, also seems to think that Churchland's seven salient dimensions of consciousness are important phenomena (see Dennett and Densmore 1999).

Claim: The word processing functions of your digital computer are the result of a virtual machine implemented by software onto the serial hardware of the computer.
Objection: The word processing functions of your digital computer are *actually* the result of a high speed CPU, random access memory, and a store of data and instructions on the fixed disc.

This diagnosis of the debate might well be accurate, but Churchland's criticism in the above quote is (finally) one worth taking seriously. In short, he claims that the seriality of consciousness is not the result of training. This is a real possibility, and also turns out to be one of the most substantial points of (genuine) dispute between Dennett and Churchland. The following sections examine this issue more closely.

Section 6: Sequentiality

6.1 Churchland's Consciousness

Notice that in the final quote of the last section Churchland claims that his recurrent networks have the capacity "for recognizing, manipulating, and generating serial structures in time" and that this means we no longer need vNvms. However, Churchland's proposal only covers, what I argue, is *one* of the explanatory functions of Dennett's proposal, viz. sequentiality. So even if he is right about this aspect of Dennett's theory, the conclusion that the need for a vNvm has been lifted is not correct, since under my reading, we are left with an explanatory surplus (viz. awareness, symbols, and control). This section then, examines Churchland's proposal for the feature of consciousness I have called 'sequentiality'. This feature remember, is for my purposes, dealt with in abstraction from the symbolic nature of (some of) the contents of consciousness, and the (sometimes) ordered character of a stream of consciousness.

The contrast case for a recurrent network is a feed-forward network, which transfers information from a sensory or input layer through the hidden unit layer and straight on to the output layer. Recurrent networks have an additional property - information processed at the hidden layer is still sent on to the output layer, but it is also sent back, via descending pathways, to re-enter the hidden layer, where it is processed together with new incoming information. Additional nodes, appropriately named 'context units' are added along these descending pathways to regulate the amount of influence that the already processed information should have on the hidden layer's activities. This feature allows networks to succeed with tasks that require new information to be processed in light of past information, or in other words, to deal with structures in time.

This feature, argues Churchland, offers an elementary form of short-term memory,[21] since recently processed data circulates back down recurrent pathways and re-enters the system with the present data. Although this provides an extremely short amount of

[21] This is one of Churchland's seven salient dimensions of consciousness.

memory, Churchland suggests "that progressively larger networks, with multi-staged recurrent loops and real-valued coding, will display a short-term memory that reaches progressively farther back into the past." (1995: 217)

In summary, Churchland contends that recurrent networks have a natural and "dramatic ability to generate complex representations with a continuously unfolding temporal dimension" (1995: 267). According to Churchland we have many recurrent loops in our brains, several distinct senses, and yet one unified consciousness.[22] So how could this unity (and sequentiality) be possible?

Churchland describes "an important system of neuronal pathways that connect almost all areas of the cerebral cortex, and subcortical areas as well, to a central area of the brain's thalamus called the *intra laminar nucleus*." This is an area of the brain which is phylogenetically very old, developing long before the cerebral hemispheres, and which "projects long axons that radiate outward to all areas of the cerebral hemispheres... (and) also receives systematic axonal projections returning from those same areas" (Churchland 1995: 215). We will go into a little detail about the presumed (and relevant) functions of the thalamus in due course, but for now it is important to note that the overall arrangement constitutes a large recurrent network embracing the whole cortex with a bottleneck in the intralaminar nuclei (Churchland 1995: 215).

The first question readers of *Consciousness Explained* might ask here, is whether this move advocates the much maligned Cartesian Theatre idea which states, more or less, that all conscious activity is only such because it passes through a single place, or over a determinate finish line? The answer, as Dennett and Densmore assert is 'no', because Churchland is not claiming that arrival of information at the intralaminar nucleus suffices to make it conscious. The intralaminar nucleus is seen as "a coordination and distribution centre" only, and importantly, the sequence in which particular contents arrive there need not have any bearing on the subjective order of those contents in conscious experience (Densmore and Dennett 1999: 759).

---

[22] This too, is one of the seven salient dimensions of consciousness.

Churchland speculates that the intralaminar nucleus and its radiating axonal projections provide some useful clues for explaining the integration of our sensory modalities in conscious experience, he writes:

> "Information from all of the sensory cortical areas is fed into the recurrent system, and it gets jointly and *collectively* represented in the coding vectors at the intralaminar nucleus and in the axonal activity radiating outward from there. The representations in that recurrent system must therefore be *polymodal* in character. This arrangement is also consistent with the familiar fact that, through oxygen depravation or anaesthetics, one can lose visual consciousness while briefly retaining, for example, auditory and somatosensory consciousness. In such a condition, we may speculate, the recurrent system ... is still functioning, but the loop that includes the visual cortex has lost function slightly ahead of the other loops." (1995: 222)

This information alone does not give us good enough reason to think that this system plays an important role in consciousness. So before drawing any conclusions there is further information about this system that we will need to consider.

Churchland discusses the results of research done by a neurophysiologist called Rodolfo Llinas, who discovered, using a technique called *magnetoencephalography* (MEG), that there exists "a steady oscillation in the level of neural activity in any area of the cortex, an oscillation of about 40 cycles per second. Moreover, the oscillations in distinct areas all stood in a constant phase relation to each other ... this phase-locked activity indicates that in some way or other they must all be parts of a common causal system" (Churchland: 1995: 219). It seems that there is some kind of connecting and coordinating system behind this phenomenon, and the best candidate we have to explain this are the large recurrent projections stemming from the intralaminar nucleus, especially because, independent research has shown that the neurons in the intralaminar nucleus have an

"intrinsic tendency, (when active), to emit bursts of activity at the required 40Hz"(Churchland 1995: 220).

The steady 40Hz oscillation is present whether one is awake or dreaming and to a lesser degree in deep sleep. However, when we are awake it is overlaid by "large non-periodic variations in the level of neural activity" which are unique to different parts of the cortex, and which are correlated with changes in the perceptual environment such as lights going off or tones being heard (Churchland 1995: 220). When we dream, the 40Hz background hum is still present, and is similarly overlain with non-periodic oscillations in the collective activity of the cortex, which quite obviously, are not correlated with changes in the perceptual environment. And finally, when we are in deep non-dreaming sleep the 40 Hz oscillation is of lower amplitude, and the neurons of the intralaminar nucleus are inactive.[23] (Churchland doesn't tell us how the oscillations could still be present even though these neurons are inactive, but it could be the case that although the 40Hz oscillation is coordinated by the intralaminar nucleus, its inactivity does not cause an immediate chaotic reaction in the cortex, but rather a slow degradation of orderliness. This however is just speculation.)

Finally, research on both humans and experimental animals has shown that damage to one side of the intralaminar nucleus produces a "hemineglect of everything having to do with the connected side of the animal's body, both sensory and motor" (1995: 220). And bilateral damage produces a "profound and irreversible coma". In conclusion the intralaminar nucleus seems to have some sort of combinatorial or assimilating effect, like a telephone exchange between all areas of the cortex, and without this coordinating and communicative device, consciousness (in Churchland's view) is just not possible. On the basis of the evidence above and his vector coding theory of representation (see Churchland 1995: 35-57), Churchland suggests that:

---

[23] These ideas form the main part of Churchland's recommendations for two more of the seven salient dimensions of consciousness: consciousness disappears in deep sleep, and, consciousness reappears (at least in a muted, disjointed manner) during dreaming.

> "a cognitive representation is an element of your current consciousness if, but only if, it is a representation – an activation vector or sequence of vectors – within the broad recurrent system. Your brain has many other representations or course, but the story just outlined entails that they are not part of your active consciousness" (1995: 223).[24]

This may be a too-simple theory of consciousness, but the far-reaching recurrent network that emanates from the intralaminar nucleus at least suggests an anatomical mechanism for the problem of sequentiality. I say this, firstly because it is undoubtedly a system which plays an important role for consciousness, secondly, it is a system resembling a recurrent network which, as we have seen, quite naturally recognise, manipulate and generate serial structures in time, and thirdly because activations occurring all over the cortex are integrated and distributed through its operation - suggesting a possible explanation for why our sequential conscious experience is a synthesis of all our sensory modalities (and more besides perhaps).

Not being a neuroscientist (or even a neuroscience student), I am in no position to adequately evaluate Churchland's theory. Instead, lets examine the implications it may have if correct. Firstly, the paradox of how we get a serial phenomenon from activations that are distributed all over the brain – one of Dennett's major motivations for proposing the virtual von Neumann machine (1991a: 214) - could conceivably have an anatomical, rather than a programmed, explanation:

> "...notice that, on my account of consciousness, the configuration of the network's synaptic weights (what Dennett wants to call its "program") plays no explanatory role in the emergence or generation of consciousness itself. The weight configuration will indeed play a decisive role in what *concepts* the creature deploys, what *features* of the world it cares about, and what *behaviours*

[24] Other theorists have similar ideas on consciousness. For example see Gerald Edelman's (1992: 117) discussion of the 'thalamacortical system', Bernard Baars' (1988) proposal, called the 'reticular thalamic activating system', and Francis Crick's (1984) 'searchlight hypothesis'. All involve similar parts of the brain and are put forward as partial theories of consciousness.

> it is capable of producing, but it plays no role in giving rise to consciousness in the first place." (1999: 766)

I will return to this contention in the next section, once we have looked into Dennett's proposals for explaining sequentiality.

6.2 Tarzan, Turing, and Sequentiality

Dennett's model of consciousness describes many coalitions of specialists competing in parallel pandemonium, creating Multiple Drafts as they perform their various activities (1991a: 135). Then: "Onto this substrate nervous system we now want to imagine building a more human mind, with something like a "stream of consciousness" capable of sustaining the sophisticated "trains of thought" on which human civilization apparently depends" (1991a: 189). To do so Dennett describes a host of self-directed stimulations, and cultural habits that lead to a fairly consistent pattern of *winners* amongst the pandemonium of competing coalitions, which in turn leads to the installation of the vNvm, which is then responsible for some drafts getting "promoted to further functional roles" (1991a: 254).[25]

> "...the overall structure of the new set of regularities, I suggest, is one of *serial chaining*, in which first one 'thing' and then another 'thing' takes place in (roughly) the same 'place'". This stream of events is entrained by a host of learned habits, of which talking to oneself is a prime example." (1991a: 221)

Dennett's argument (heavily simplified) seems have the following structure: we have a physical system that operates in parallel, and we know this system to be responsible for a sequential phenomenon – conscious experience. Since we inherited the parallel system it must be through learning that it comes to exhibit a sequential phenomenon. Even if this is a true aetiology we still require an understanding of how the (now meme-infected) brain

---

[25] This concept (of 'promotion') will be discussed in section (7.6).

comes to behave such that it can support this sequential phenomenon. Or rather, we need to know how certain features of enculturation affect the broad/global processing dynamics of the brain. Does Dennett's theory provide this?

Lets proceed through an analysis of two possible criticisms of Dennett's position on sequentiality: Although we may concede that there are important cultural habits that, once inculcated, lead to an *ordered and purposeful* stream of consciousness. It is possible that a less focused, less disciplined, non-symbolic, but still sequential stream of consciousness could exist in, for example, infant children, feral children, or chimpanzees, and that enculturation imposes augmenting and enhancing regularities on a system that already exhibits a sequential conscious experience (call this criticism 1). In addition, one might argue that even if Dennett is right and no such stream exists without enculturation, he has explained how the stream *comes to exist*, and not how the (now altered) PDP brain works such that it supports this stream (call this criticism 2).

Lets begin with criticism 2. Rumelhart et al (1986) describe a class of neural networks, relaxation networks, which are (more or less) functionally identical with Dennett's pandemonium idea:

> "Since we assume that the system is moving toward a maximum goodness solution every time a new input comes into the system, the system operates in the following way. An input enters the system, the system relaxes to accommodate the new input. The system approaches a relatively stable state which represents the interpretation of the input by the system. The system then occupies this state until the stimulus conditions change. When a new input arrives, the system relaxes to a new state. (...) Since it is patterns of activation over a set of neurons that are the relevant representational format and since a set of units can only contain one pattern at a time, there is an enforced seriality in what can be represented. A given set of units can, however, be seen as representing a sequence of events." (Rumelhart et al 1986: 38)

In general, there is little difference between 'winners' emerging from Dennett's pandemonium model, and relaxation networks moving towards 'a maximum goodness solution', since both respond to inputs by a process of arranging, or striving, or competing, or relaxing, until the system *settles* into a *best fit*, (or the strongest competitors win). A succession of such *relaxations* necessarily reveals a serial component, and thus an interesting perspective on the relation between parallel and serial processing. What emerges when relaxation networks are observed over time is a sequence of stable states (with re-organizing in response to new inputs in-between).[26] How does this idea mesh with the *virtual* seriality of the vNvm?

Lets look at the problems we face: Dennett is perhaps a bit vague about the relation between his Multiple Drafts model, his vNvm theory, and the sequentiality of consciousness. It seems almost certain that Dennett's Multiple Drafts model is supposed to account for sequentiality: "These (multiple drafts) yield, over the course of time, something rather like a narrative stream or sequence..." (1991a: 135). Yet Dennett and Densmore (1999) stress that the "only feature of von Neumann architecture that Dennett imputes to the "von Neumannesque" virtual machine he is discussing is its seriality (...)" (1999: 748). In addition Dennett tells us: "the seriality of this machine (its "von Neumannesque" character) is not a 'hard-wired' design feature, but rather the upshot of a succession of coalitions of theses specialists" (1991a: 258). This too is strange, because Dennett does not say that the instantiation of his Multiple Drafts model is the result of enculturation (i.e. it seems that the Multiple Drafts model does track our 'hard-wired design features' 1991a: 254). (Much here hangs on our interpretation of 'upshot' too; a vague term, in my opinion, for the relationship between the two main pillars of Dennett's theory of consciousness.) As we will see below, I think the way forward is to take the term 'seriality' (in the context of vNvms) to mean quite a lot more than mere sequentiality.

[26] It is tempting to speculate that these 'relaxations' (if they occur at all in real brains) may occur at the 40Hz oscillations we discussed in the previous section, but I have not been able to find anything to corroborate this.

Rumelhart et al's relaxation networks provide, in my view, a preliminary and partial suggestion of how the Multiple Drafts model might be implemented in real 'brain networks'. Lets say this is the case. We could then describe a system that supported a pre-existing capacity for sequential consciousness (much as the Multiple Drafts model does), which could then be vastly augmented and transformed (by enculturation), into a system whose sequences *become* an ordered, purposeful, stream of symbols (or rather, *virtual symbols*), and this could be what Dennett means by 'seriality'.

This may be the correct interpretation of Dennett's ideas, but the association between seriality and vNvms (cited above), and between Multiple Drafts and "a narrative sequence" (1991a: 135) strikes me as ambiguous. If the above interpretation is correct, then Dennett must deny that it is special training that leads to the (bare) sequential phenomenon, and further, argue that what he means by 'seriality', is not (my) sequentiality, but rather the *ordered, symbolic* sequentiality of an educated adult human's consciousness. If by saying that "the overall structure of the new set of regularities ... is one of *serial chaining*", Dennett is referring to the sequential nature of consciousness (period), and not the ordered and purposeful nature of that phenomenon, then it would seem that he has accounted for the same phenomenon twice. I doubt that this is the case and so will persist with the prior interpretation.

In light of criticism 2 then, Dennett could claim that his Multiple Drafts model provides (at least) a suggestion of how sequentiality could work in brains, and the vNvm idea refers to the ways in which this (basic feature) is enhanced such that it can support a *human* stream of consciousness. Here then, is where criticism 1 emerges, since if Dennett makes the move suggested above in response to criticism 2, he must then argue for why the relaxation networks (or multiple drafts) of infants, feral children or chimpanzees do not deliver this phenomenon – viz. a sequential (albeit impoverished) conscious experience. Alternatively (and correctly I think), Dennett could concede that sequentiality is (rather, could be) a feature present in, for example a chimpanzee, but that this (whatever it is) is not consciousness. Indeed, Dennett and Densmore claim that it is likely

that human consciousness is so "different from that of any other species that to call the other varieties consciousness is to court confusion" (1999: 759). And more precisely:

> "In order to be conscious – in order to be the sort of thing it is like something to be – it is necessary to have a certain sort of informational organization that endows that thing with a wide set of cognitive powers (such as the powers of reflection, and re-representation). This sort of internal organization does not come automatically with so-called sentience. It is not the birthright of mammals, or warm-blooded creatures, or vertebrates; it is not even the birthright of human beings. It is an organization that is swiftly achieved in one species, ours, and in no other. Other species no doubt achieve *somewhat similar* organizations, but the differences are so great that most of the speculative translations of imagination from our case to theirs *make no sense*." (1995: 347)

The idea of 're-representation' will occupy us in coming sections. For now, recall that Churchland considers it a virtue that on his account that "creatures with inchoate or still poorly developed conceptual frameworks – creatures such as human infants and non-human animals – still can and must be conscious in the same ways that we adult humans are. Their concerns may be narrow and their comprehension dim, but when they wake up in the morning they, too, are conscious" (1999: 766). Thus, Dennett might argue that although Churchland's theory is applicable to a number of (higher) creatures, his theory does not in fact explain consciousness (as we *should* understand it). This all sounds very confusing, so lets try to get things a little clearer.

I think that the preceding sections positively *beg* for some kind of distinction. Perhaps we should talk of two kinds of consciousness, say, of awareness and consciousness, sentience and consciousness, or as Gerald Edelman does, of "primary consciousness" (1992: 117-23) and "higher order consciousness" (1992: 124-36). Or even, as I prefer: Tarzan[27] consciousness and Turing Consciousness. The latter would only be available to

[27] An examiner rightly pointed out that in the novels (before the arrival of Jane) Tarzan was fluent in both Ape and Elephant. This makes my choice of label rather inapt since I am referring to a consciousness that is

(enculturated) humans; characterized by a thinking subject with a stream of awareness of his or her own acts, thoughts, emotions, etc; with competencies like serial reasoning, consideration of past and future events, symbolic thought, language comprehension, visual imagination, and so forth. The latter, Tarzan consciousness, is arguably (on Dennett's account) not something we should call 'consciousness' at all, but some other thing (perhaps unconscious integration, real-time responsiveness, behavioural and environmental awareness, whatever.) The latter is also, I might add, close to the way Churchland sees the problem of consciousness: more or less, the problem of accounting for what happens when we wake-up from a sleep. Or perhaps more fairly: a unified, multi-modal stream of environmental, sensory and bodily integration and coordination.

To summarize: Although surely too idealized, the distinction above allows for the idea that (re consciousness) Churchland is tracking a different phenomenon to Dennett. Churchland's claim that "the human brain does not have to be carefully *programmed* in order to exhibit the basic dynamical features of short-term memory, steerable attention, plastic interpretation, sensory-free activity (both day-dreaming and night-dreaming), quiescence (deep sleep), and polymodal integration." And that: "These basic features arise automatically from the basic physical structure of the system, independently of the details or the maturity of its conceptual development" (1999: 766) may be correct for Tarzan consciousness, but inadequate to account for Turing consciousness.

The features of Turing consciousness, (or the von Neumannesque features) could then be precisely those which Dennett tells us are invisible to neuroanatomical scrutiny (1991a: 219), those which require programming, and those which demand to be accounted for at the virtual machine level of explanation. Explaining consciousness then, would be very analogous to the following:

Suppose I purchase an unassembled bicycle. Following the assembly instructions I fit all the parts together and finish with a working bicycle. We might now ask: Is the

---

completely free from public language and socio-cultural influence. Readers are advised to imagine my sort of Tarzan in what follows.

functionality of the bicycle best explained by the fashioning of the individual parts in the factory, or by my diligent following of the assembly instructions. Quite obviously this is an inane question. Without *both* the original parts *and* my construction there would be no bicycling whatsoever. Thus, the question above tracks a false dichotomy and forces a choice that we should never have to make. The constituent bicycle parts, and their subsequent assembly, are both relevant and important to a complete explanation of the bicycle's functionality.

In a similar way, both Dennett and Churchland's theories of consciousness may be relevant and important to a complete explanation of consciousness. In other words the capacities of our neural architecture (visible to neuroscience) and our developmental enhancements (best captured from a design stance, program or virtual machine level) are both relevant and important to a comprehensive explanation of human consciousness. Churchland's mechanistic theory may get us very far in accounting for sequentiality and the consciousness (if we should call it that) of non-human animals, but the 'final' *process* explanations for an untrained, un-meme-infected consciousness may well be indistinguishable from the enhanced 'higher order' varieties, and in order to have explained human consciousness, we need an account of the differences underlying these advanced competencies and experiences. In summary, if Dennett is correct in claiming that the signature characteristics of adult human consciousness are the result of broad, epigenetically based, patterns of neural activity, visible only from a high-level of analysis, then Churchland's theory, (quite unfortunately), will not be equipped to track the difference between Tarzan and Turing.

6.3 Re-representation

We are now in a position to examine some crucial aspects of human/Turing/higher-order consciousness. Gerald Edelman cryptically describes the distinctive feature of this phenomenon as a capacity to be "conscious of being conscious" (1992: 112), but what does this circular sounding phrase mean? Further, and more to the point, what is it about the installation of a vNvm that makes human consciousness *happen*?

As we now know, Dennett describes the analogy with the virtual machines of computer science as providing a useful perspective on the phenomenon of human consciousness, he notes: "Computers were originally just supposed to be number crunchers, but now their number crunching has been harnessed in a thousand imaginative ways to create new virtual machines, such as video games and word-processors" (1991a: 225). And similarly: "Our brains (...) weren't designed (except for some very recent peripheral organs) for word processing, but now a large portion – perhaps even the lions share – of the activity that takes place in adult human brains is involved in a sort of word-processing: speech production and comprehension, and the serial rehearsal and rearrangement of linguistic items, or better, their neural surrogates." (Dennett 1991a: 225)

Dennett predicts that readers will object: a von Neumann machine is utterly unconscious; "why should implementing it – or something like it – be any more conscious?" (1991a: 225) Dennett's answer to this: "The von Neumann machine, by being wired up from the outset that way, with maximally efficient informational links, didn't have to become the object of its own elaborate perceptual systems. The workings of the Joycean machine, on the other hand, are just as "visible" and "audible" to it as any of the things in the external world that it is designed to perceive – for the simple reason that they have much the same perceptual machinery focused on them" (1991a: 226).

Although not ostensibly about consciousness the following is perhaps a clearer characterization of this idea: "The human capacity to treat one's own states, reflexively, as the objects of further consideration – perception and thought – greatly enhances our powers of understanding. Arguably, the sort of swift, sensitive self-redesign it permits is the basis for our undeniable cognitive superiority over all other organisms." (1992: 77)

Recall that in the last section Dennett referred to a "certain sort of informational organization" which was fundamental to consciousness. This arrangement allowed for "a wide set of cognitive powers (such as the powers of reflection, and re-representation)".

The following section explains one hypothesis that appears to provide just the kind of informational organization we need.

Section 7: Symbols and Consciousness

7.1 Flexibly Deployable Contents

The theory of representational redescription (hereafter RR) is originally the brainchild of Annette Karmiloff-Smith (1985; 1986; 1987). However, a paper co-authored by Andy Clark and Karmiloff-Smith (hereafter: CKS 1994) will be most fruitful for our purposes.[28] RR is a theory that parallels many aspects of the virtual machine hypothesis. As section (7.6) will suggest, RR has much to do with the development of consciousness, and also blurs the *line* between serial and parallel processing. For now it is important to note two other similarities: (1) the Representational Redescription model, like the vNvm idea, describes the general features of human cognitive life that mark a major division in the natural order between humans and all other cognitive creatures. (2) The main focus of the paper is on what CKS see "as the basic psychological and computational issue, *viz.* how to achieve flexibility, manipulability, and transportability within a broadly connectionist setting." (CKS 1994: 513) In a similar way some of the benefits the vNvm supplies are precisely those of flexibility, manipulability and transportability (the essential characteristics of symbols). It would seem worthwhile then, to establish some understanding of the hypothesis, and investigate the main differences between RR, and the installation of a vNvm. In a nutshell, representational redescription is:

> "an internal organization which is geared to the repeated redescription of its own stored knowledge. This organization is one in which information already stored in an organism's special purpose responses to the environment is subsequently made available, by the RR process, to serve a much wider variety of ends. Thus knowledge that is initially embedded in special-purpose effective procedures subsequently becomes a data structure available to other parts of the system." (CKS 1994: 488)

[28] This is because it places RR firmly in the context of connectionist research, and was reviewed by Dennett (1994c) forming a good basis for discussion in section (7.3 and 7.4).

Karmiloff-Smith (1985; 1986; 1987) has collected vast quantities of evidence in support of the RR model from carefully designed experiments. I will not discuss many of these but to supply some idea: most often the experiments involved an analysis of performance on a set task, by children from different stages of development. The tasks usually involved 'cross-domain' use of knowledge such as drawing a 'funny man', where children would be asked to modify their (usually stereotyped) picture of a man, by adding features from other areas (like wings or horns). Children below a certain stage of development could not perform the task successfully. The hypothesis CKS endorse is that the child's knowledge of drawing a man is initially highly inflexible and task-specific, (as is knowledge from other domains), and this results in the impossibility of the combinatorial task (CKS 1994: 495-503). I will supply further details as we proceed.

7.2 Inter-Network Knowledge Transfer

The RR model has some interesting parallels with connectionist research and pursuing this avenue will help to clarify the details of the RR model. A major stumbling block for connectionist research is "the inter-network knowledge transfer problem". For example, the cluster analysis of NETtalk (Sejnowski and Rosenberg 1986) revealed that its activation space had been partitioned so as to treat a,e,i,o,u as similar. However NETtalk could not make use of the abstractions 'vowel' or 'consonant', it could not so much as pick a vowel from a string of letters. Even though, we external theorists can *observe* (through cluster analysis) that the network treats vowels and consonants differently, *as if* it has some sort of knowledge of the practical difference between the two. CKS nicely describe this as "knowledge *in* the system but not knowledge *to* the system". Overcoming the inter-network knowledge transfer problem then, involves exploiting the results of successful learning in one network immediately in another, thereby eliminating the need for retraining on that aspect of the task. As CKS state: "in the absence of such transportable knowledge, networks are unable to progress beyond the limited task specific use of their knowledge." (CKS 1994: 490)

In the context of NETtalk this progression would involve being able to exploit its knowledge of the difference between vowels and consonants, without losing its original functionality. The inter-network knowledge transfer problem derives essentially from the fact that 'first-order networks' (networks like NETtalk) contain intricately interwoven aspects of knowledge in a single knowledge structure, and this poses problems for flexibility and generality. A network may be capable of making generalizations when exposed to a series of inputs, but these generalizations are not represented such that they can be exploited for use in other tasks.

Just like first order networks, CKS suggest that human learning begins by forming just this kind of interwoven knowledge. But humans are "internally driven to go on to form a series of further representations which allow us to manipulate and exploit our knowledge in increasingly flexible, and mutually independent, ways" (CKS 1994: 494-495). The process of RR is postulated as a system whereby specific aspects of an originally task-specific, and heavily interwoven knowledge structure can be extracted and exploited by other sub-systems, whilst leaving the original network intact. Essentially RR is an internal process of "representational change whereby the mind enriches itself from within by re-representing the knowledge that it has already represented." (CKS 1994: 495)

7.3 Innately Redescribers

CKS are careful to stress the incremental stages of the re-description process that they postulate for humans, (in opposition to views which might characterize development as merely involving a "dichotomy between implicit and explicit linguistically-encoded representations"). At the first stage, called implicit, or level-1 representations, knowledge is represented and activated in response to external stimuli, but it is not yet available for use by any other part of the system. More exactly: "A procedure for producing a particular output is available as a whole to other processes, but its component parts (i.e. the knowledge embedded in the procedure) are not." (CKS 1994: 495). Level-1 describes the base level from which the stages of RR progress, and corresponds to an early phase of cognitive development in real children. The first level of redescription is termed E1

where condensed representations are more explicitly defined. However, E1 representations are not yet available for conscious access or verbal report – this requires further levels of redescription, E2, E3, etc. As redescriptions are redescribed, representations become progressively more flexible and accessible, or rather, much more like symbols (CKS 1994: 496).

CKS conclude that:

> "developmental studies provide clear evidence that such a progression occurs in human cognition and that it involves a (conservative) series of phases in which constraints on the use of acquired knowledge are gradually relaxed. The crucial question – by what concrete computational mechanisms may such relaxation be achieved? – remains unresolved" (1994: 513).

The problem of a 'concrete computational mechanism' is one that can perhaps be investigated by attempting to build more sophisticated connectionist networks. As CKS point out, if networks are to become suitable models of human RR they must be able to treat their own representations as objects for their own manipulation, to do this independently of prompting by continued training inputs, and without losing the functionality of the original networks, such that they can "form new structured representations of their own knowledge which can be manipulated, recombined and accessed by other computational processes" (CKS 1994: 509). Quite obviously, not just any network architecture would deliver this, but it could be the case that in addition to this architecture (whatever it is), specially tailored training is required to deliver the requisite competency.

In the human case, the issue above concerns the degree to which humans are innately 'redescribers', as opposed to acquiring this capacity purely through interacting with the socio-cultural/linguistic environment. Something like the former view is held by CKS (although they believe this innate ability is *enhanced* by the socio-cultural/linguistic environment), whereas Dennett endorses a version of the latter position. CKS write:

> "the RR model postulates that external stimuli, for example new input, failure of a procedure to bring about a certain goal, communicative pressures from others, etc., are not a necessary condition for representational redescription to take place. In the main the RR process is generated from within. It does not depend on information currently coming into the system for processing." (CKS 1994: 502)

As mentioned, how RR is 'generated from within' is still in doubt, but the compelling behavioural evidence Karmiloff-Smith has collected implies that (somehow) such a process must be taking place. This evidence suggests an innate disposition towards RR in two ways, firstly: "the careful study of developmental data shows that the increased behavioural flexibility often pre-dates the subject's ability to formulate or understand linguistically encoded expressions about the components and rules governing the domain." CKS go on to state, "that the possibility of such expressions is an effect of a prior process of representational redescription, not a substitute for it." Secondly, "developmental data suggest that frequently the ability to re-configure the linguistic description of a task is not of any help" (CKS 1994: 505). Thus, although the RR process "is indeed *sometimes* generated by externally provoked stimuli, it is a predominantly endogenous process hypothesized to differentiate human cognition from that of most (all?) other species" (CKS 1994: 503).

7.4 Labels Against Innateness

CKS describe theories that make the distinctive cognitive abilities of humans dependent on ability with a natural language (like Dennett's theory), as "getting the cart before the horse" (1994c: 505). They contend that language is "just one more level of re-description", although they grant that it provides "rich manipulability" and a powerful form of cultural transmission (1994c: 505).

Dennett (1994c) in his response to CKS asks: "Does the advanced use of language depend on an RR capacity that develops, but is "specified in innate predispositions" or

does the RR capacity that develops depend to an important degree on what the child *acquires* when the pre-designed structures of natural language are moved from the child's enveloping culture into its brain?" (1994c: 541) And in answering this question Dennett suggests: "the capacity of a system to engage in representational redescription really does depend on that systems capacity – not yet fully developed but in the process of development – to master and use a natural language" (1994c: 541). Recall that Dennett argues that the human capacity for language use is supported by, a combination of the plasticity of the brain, the novel use of inherited competencies (which were *designed* for other abilities), and exposure to the pre-developed system of a natural language. And so Dennett highlights the possibility that exposure to language (somehow) initiates RR, and that the now 'up and running' RR process then leads to the *advanced* use of language. The problem is to account for the 'somehow'. Or rather, to explain how this 'bootstrapping' process gets initiated. Lets look briefly at some of Dennett's ideas:

Dennett concedes that a child in the process of gradually equipping itself with new competencies does not always "directly involve the child's using natural language explicitly directed to the task at hand". But he nevertheless suggests that what is going on during this process "might nevertheless depend on the natural language competence the child is acquiring" (1994c: 542).

To make the view a little clearer Dennett pursues an example of a child who, when reaching towards a hot plate in the kitchen, hears words like, "hot!" and, "don't touch the stove", from her mother. Dennett describes this as "conjuring up a situation-type" and importantly, "not just a situation in which a specific prohibition is typically encountered but also a situation in which a certain auditory rehearsal is encountered" (Dennett 1994c: 542). Following this Dennett suggests that the child would rehearse the heard words quite spontaneously: "after all children are taken with the habit of rehearsing words they have just heard,[29] in and out of context, building up recognition links and association paths

[29] Readers may be given to ask: on what does *this* habit depend? Or speculate that this habit sounds rather like the early phases of RR postulated by Karmiloff-Smith. I consider this to be a possibility, but cannot pursue it fruitfully without a more detailed account of both author's positions on early childhood development – something for which there is no space in the current work.

between the auditory properties and concurrent sensory properties, internal states, and so forth." This "semi-understood self-commentary" is according to Dennett the beginnings of a habit of labelling; a habit that is created before the labels had to be (even partially) understood. 'Semi-understood self-commentary', says Dennett, could be the origin of the practice of deliberate labelling, which could lead to the more efficient practice of dropping all or most of the auditory and articulatory associations and just "relying on the rest of the associations to do the anchoring". The child can thus abandon out-load speaking and "create private, unvoiced neologisms" as labels for features of its own activities (1994c: 543).

Thus, Dennett's argument is that the labelling function of words could be an epigenetic, and initially exogenous, feature allowing for representational redescription without an inner mechanism for just that purpose.

> "...we can take a linguistic object as a *found object* (even if we have somehow blundered into making it ourselves, rather than hearing it from someone else), and store it away for further consideration, "offline". This depends on there being a detachable guise for the label, something that is independent of meaning. Once we have created labels, and the habit of "attaching" them to experienced circumstances, we have created a new class of objects that can themselves become the objects of all the pattern-recognition machinery, association building machinery, and so forth." (Dennett 1994c: 544)

Clark echoes this idea:

> "The claim is thus that associating a perceptually simple, stable, external item (such as a word) with an idea, concept or piece of knowledge effectively freezes the concept into a sort of cognitive building block – an item that can then be treated as a simple base-line feature for future episodes of thought, learning and search." (1998: 174)

Recall that CKS argue that language use is not necessary for RR because increased behavioural flexibility often occurs before a child's ability to "formulate or understand linguistically encoded expressions about the components and rules governing the domain" (1994c: 505). Dennett argues that the child, even without 'understanding', could nonetheless achieve flexibility through prior semi-understood exposure to labels. I present this to suggest the coherence of Dennett's perspective and not in hope of deciding the issue either way. Whatever its origins, RR seems to be crucial to the development of language use and the type of symbolic consciousness we enjoy, and, as I discuss below (section 7.6), RR is perhaps crucial to the development of conscious awareness itself.

7.5 Another Virtual Disagreement

This short interlude highlights Andy Clark's delicate criticism of Dennett's vNvm idea outside of the context of RR. However, it is not at all clear that (in this context) Clark's position is all that different from the one he criticises. Clark sees language as an extremely well designed 'tool' which we humans are well designed to use: "Where Dennett sees public language a effecting a profound but subtle reorganization of the brain itself, I am inclined to see it as in essence an external resource which compliments – but does not profoundly alter – the brain's own basic modes of representation and computation" (1998: 167). Clark does not want to treat language as wholly external, but claims that the changes that occur in the brain when a language is acquired are relatively superficial, and that genetically inherited capacities (like pattern-completion) are responsible for the efficient use of this tool (1997: 198).

Both Dennett and Clark believe that some kind of internal effects must take place when a language is learnt. This much is obvious from common experiences like mentally rehearsing sentences in our heads, and speaking (without constantly referring to external props). Where their views are *seemingly* opposed is the degree to which the brains basic functionality gets altered by language. I think Clark is mistaken to take the virtual machine idea as something that profoundly alters the *basic* ways in which the brain *represents and computes* (1997: 198; 1998: 167). Dennett's position is not that the

parallel-distributed system is substituted by the virtual machine; the basic inherited *capacities* of the brain are not replaced, but vastly enhanced, and used in new and innovative ways. Thus, Dennett's claim is much like Clark's; namely, that exposure to language leads to *superficial* changes at the neural processing level (the basic modes of representation and computation), but that these changes nonetheless amount (at a higher level of analysis) to the installation of a system of moveable symbols.

7.6 Consciousness Redescribed

Up until now I have set aside some aspects of RR that make it a (partial) theory of consciousness. Considering this possibility augments the speculative discussion of 're-representation' in section (6.3):

> "consciousness, on such a model, would be a side effect of the drive to form ever more abstract and structural representations of one's own stored knowledge. (This link between multiple levels of redescription and consciousness has always been at the heart of the RR model.) (...) The role of redescription is, then, twofold. It is the means whereby we achieve higher orders of flexibility and (more speculatively) the root of our conscious awareness of structured mental states." (CKS 1994: 513)

Contrast this with the following from Dennett's model of consciousness:

> "Most of these fragmentary drafts of "narrative" play short-lived roles in the modulation of current activity but some get promoted to further functional roles, in swift succession, by the activity of a virtual machine in the brain" (1991a: 254).

In our now developed context, the 'innateness discussion' above could be re-stated as one concerning the question of how drafts are 'promoted' before the complete installation of the virtual machine. More interesting though, is the way that the RR model provides a

more detailed way of explaining Dennett's 'promotion'. In other words one primary function of the vNvm seems to be almost indistinguishable from representational redescription. What this shows, I suggest, is that finding ways in which RR is implemented in brains would go along way towards explaining consciousness, and also towards showing how brains come to exhibit a competence for two features of the vNvm from section (4.3), namely conscious awareness, and symbols. In sum, the 'informational organization' that Dennett argues is necessary for consciousness could be an organization which allows for representational redescription, or rather, one that *results from* representational redescription.

Interestingly, there remains a further similarity between Dennett's vNvm idea and RR that relates back to our earlier themes:

> "it is important to note that the RR model raises two potentially distinct computational issues. The first is: how does redescription occur? That is, what *kind* of internally driven computational process yields the higher-level redescriptions? The second is, in what does the redescription consist? That is, what is the character of the higher-level representational formats? There is thus a question about *process* and a question about *product*. And it is at least conceivable that these might cross-classify the connectionist/classicist issue. Thus we might have a connectionist developmental process whose product is a symbolic representational language" (CKS 1994: 506).

The link with Dennett's theory should be fairly clear from our discussions in section (2). In short, (part of) the products of the installation of a vNvm, are the same as the products resulting from progress through the levels of representational redescription (namely, a capacity for a flexibly deployable, symbolic, representational language) - both theories argue that despite the nature of this product, it can be achieved by a brain whose *processes* are broadly connectionist. As I have highlighted, the major difference between the two models is their different perspectives on the innateness of RR. I cannot hope to supply an answer to this question here. Although it is an empirical question it seems to be

very hard to figure out what kind of investigation or experiment would finally settle the issue. A useful task for some future occasion would be to formulate some experiments that could help fulfil this objective, my task now however, is to briefly discuss our final feature of vNvms: control.

Section 8: Control

8.1 Model Captains

This final section looks at the last feature on the list from section (4.3): the ordered or controlled nature of consciousness. I will present some ideas that I think are highly compatible with both Dennett's theory and the RR model.

The abundance of simultaneously active specialists postulated in Dennett's Multiple Drafts model presents considerable problems of self-control. Thus one of the primary functions of the vNvm is to "adjudicate disputes, smooth out transitions between regimes, and prevent untimely *coups d'état* by marshalling the 'right' forces" (1991a: 277). Dennett tells us that various mnemonic tricks, self-manipulations, and rehearsals lead to increasing degrees of control over the pandemonium of specialists. However, we are left without much idea of the kind of organizational arrangement that these strategies might produce, or even, what kind of organization would lead to the capacity for those strategies in the first place. I don't pretend to have an adequate answer for this but will present some ideas from Rumelhart et al (1986) that suggest a possible, partial, proposal to fill this gap.

Rumelhart et al point out that the problem with relaxation networks (see section 6.2) is that the system cannot change without 'external prodding' (1986: 39). This would leave us with a model where, until the world changes, there would be no change in the contents of consciousness. However, this is limited, because as Rumelhart et al point out: (1) the environment is always changing, and (2), an interpretation leads to an action - an action changes the environment and this feeds back into the system, leading to another interpretation and another action (1986: 40). But as the authors point out this is too passive; we need a model to account for the kind of control that allows us to act in the world, and not just be 'prodded' by it.

> "Suppose, for arguments sake, that the system is broken into two pieces – two sets of units. One piece is the one that we have been discussing, in that it receives inputs and relaxes to an appropriate state that includes a specification of an appropriate action which will, in turn, change the inputs to the system. The other piece of the system is similar in nature, except it is a "model" of the world on which we are acting. This consists of a relaxation network which takes as input some specification of the actions we intend to carry out and produces an interpretation of "what would happen if we did that." Part of this specification would be expected to be a specification of what the new stimulus conditions would be like. Thus, one network takes inputs from the world and produces actions; the other takes actions and predicts how the input would change in response" (1986: 40-1).

The second network is seen as a 'mental model' of the world events. Suppose then that events in the world did not happen – "it would be possible to take the output of the mental model and replace the stimulus inputs from the world with inputs from the model of the world" (1986: 42). This Rumelhart et al tell us, would be to "run a mental simulation" (1986: 42). This ability can serve as an internal control structure[30] since a mental model influences other parts of the system in the same way as the external world:

> "This mental model would allow us to perform actions entirely internally and to judge the consequences of our actions, interpret them, and draw conclusions based on them. In other words, we can, it would seem, build an internal control system based on the interaction between these two modules of the system." (1986: 42)

And true to their word, Rumelhart et al built just such a system: "a simple two module model of tic-tac-toe which carries out exactly this process, and can thereby "imagine

[30] Adding some corroboration to this idea is Andy Clark's description of an increasing recognition in neuroscience of neuronal control structures; which are "neural circuits, structures, or processes whose primary role is to modulate the activity of other neural circuits, structures or processes" (1997: 136).

playing tic-tac-toe" (1986: 42). The two modules involved were both relaxation networks connected together. The two networks were very similar: The first is a system which takes as input, a pattern representing the board of a game of tic-tac-toe, and relaxes into a 'solution state' which prescribes an appropriate response. The second network is nearly identical – it takes as input a board position and a move, and settles to a prediction of the opponents responding move. "In short, it is a 'mental model' of the opponent. When the output of the first is fed, as input, to the second, and the output of the second is fed, as input, to the first, the two networks can simulate a game of tic-tac-toe" (1986: 49).

This idea is naturally applicable to many other domains. Perhaps most importantly, the idea of a mental model as a control structure could be just the kind of thing Dennett is tracking with one of his 'strategies of self-stimulation': talking to yourself. With a mental model of another individual we can imagine having a conversation with someone else (Rumelhart et al 1986: 43). Mental models could be related to RR in the following way: the general thrust of the RR model is that we need a model of the mind/brain that enriches itself from within by re-representing the knowledge it has already represented. Perhaps, something like the mental model idea is one mode of such enrichment. Knowledge of the world that has already been represented in one part of the system whilst performing some task could become available through RR to other parts of the system, and thus play the role of a mental model, thus augmenting the capacities of the system. It is also compatible with Dennett's idea of a 'virtual captain' where controlling functions shift from one coalition of specialists to another (1991a: 228). These virtual captains may also turn out to be mental models of world events, or parts of the brain that also serve the purpose of representing the world during action. This brief treatment of the 'control problem' is all I have space for in this work, but the general idea described above, of some parts of the brain either alternatively, or simultaneously, serving to control a behaviour and the activity of some other part of the brain is one that may go a long way towards answering it.

8.2 Serial Reasoning Out in the World

In this final section I want to pick up on some of the themes of the last section and show how our capacity for formal reasoning appears to have its origins (partly) in our ability to use the environment to make cognitive tasks easier and quicker.

Rumelhart et al argue that the reason we succeed in solving logical problems is not so much through the use of logic, "but by making the problems we wish to solve conform to problems we are good at solving" (1986: 44). This much is compatible with Dennett's picture of the more recently developed (or higher) human abilities being the result of using our natural abilities in innovative new ways (1991a: 190, 201, 209, 218).
Rumelhart et al claim that three abilities allow us to accomplish formal reasoning: (1) we have a natural ability for pattern matching; (2) we are good at modelling the world; and (3) we are good at manipulating the environment (1986: 44). Lets take a look at what this means:

> "We are good at "perceiving" answers to problems. Unfortunately, this is not a universal mechanism for solving problems and thinking, but as we become more expert, we become better at reducing problem domains to pattern-matching tasks (of the kind best accomplished by PDP models). Thus chess experts can look at a chess board and "see" the correct move. This, we assume, is a problem strictly analogous to the problem of perceiving anything. It is not an easy problem, but it is one that humans are especially good at. It has proven to be extraordinarily difficult to duplicate this ability with a conventional symbol-processing machine" (1986: 45).

The authors go on to note that not all problems can be solved in this way (by immediately seeing the solution). Few of us can look at a three-digit multiplication problem and see the answer (1986: 45). They suggest that our ability to manipulate the environment, and more specifically, the fact that we can write the problem down in a more accessible format, helps to convert such a task into one we are naturally quite good at: "Each cycle

of this operation involves first creating a representation through manipulation of the environment, then a processing of this (actual physical) representation by means of our well-tuned perceptual apparatus leading to a further modification of this representation. By doing this we reduce a very abstract conceptual problem to a series of operations that are very concrete and at which we can become very good" (1986: 45-6).

Interestingly, Rumelhart et al describe this as "*real* symbol processing" and, "we are beginning to think, the primary symbol processing that we are able to do" (1986: 46). This makes the environment a crucial extension of our minds - a theme well argued by Andy Clark (1997). Now returning to the theme of the last section: Not only can we manipulate the environment and then process it, "we can also learn to internalise the representations we create, "imagine" them, and then process these imagined representations – just as if they were external" (1986: 46). We can thus imagine writing down a multiplication problem and imagine multiplying the numbers together, and if the problem is simple enough, we can solve it in our imagination. Rumelhart et al believe that the ability to do the problem in our imagination is derivative from our ability to do it physically (i.e. with external representational tools). The 'control problem' of previous section can also be addressed in an analogous way: We can be instructed by someone else to behave in a particular way and this can be viewed as responding to some environmental event. We can recall this event and tell ourselves what to do: "we have in this way internalised the instruction (...) We believe the process of following instructions is essentially the same whether we have told ourselves or have been told what to do. Thus, even here, we have a kind or internalisation of an external representational format (i.e., language)" (Rumelhart et al 1986: 47).

In summary, our ability to perform serial reasoning *in consciousness* seems to be reliant (at least initially) on the more frequent use of our specially tailored environment (especially its linguistic aspects) as an essential partner in our cognitive pursuits. As Clark puts it: "the biological brain is fantastically empowered by some of its strangest and most recent creations: words in the air, symbols on the printed page" (1997: 218).

Summary and Conclusion

In section (6.2) I considered the possibility that consciousness is a word we should reserve for socially situated humans, and the similar idea of developing two very different concepts of consciousness. The terminological issue does not matter much, but it does seem fairly safe to claim that the inner life of frogs, dogs, chimpanzees and Tarzan-like humans (like, human ancestors from more than 10000 years ago, feral children, etc) is significantly different to that of you, dear reader, Alan Turing, and myself. I wondered whether some of the former, although not having a very complex inner life, might nevertheless have a sequential experience of the world they live in. I left this question fairly open, but in the following, final, short, and speculative hypothesis, I would like to look at how we might, one day, be able to fully explain this difference in awareness.

First, we would need to find some prerequisites (at a neurological level) for representation to occur; something perhaps like the three put forward by Dan Lloyd (1989: 64). Here, representation occurs only when there is a (1) conjunction between at least two events in a system with multiple channels (like "a state change in an and-gate or 'majority rule' threshold device" (1989: 64)), where (2) the multiple channel device is uniquely situated such that there is a set of single events sufficient for its activation, and where (3) the resulting event has the capacity to cause either another representation or a behavioural event (1989:64). If this was the case we could speculate, as Lloyd does, that the main difference between consciousness for humans and for lower creatures, is (very crudely) that humans come to have frequent instances (in their brains) where the above conditions are met, whereas lower creatures have far fewer such instances. In sum, Lloyd argues that awareness should be equated with active representation, and thus that the more active representations occurring within a creature the more aware that creature is (1989: 192). There is much more to Lloyd's theory than I have presented here, and perhaps a future work could investigate the ways in which his ideas (in detail) could work with Dennett's theory of consciousness and representational redescription. Nevertheless, Lloyd's theory may be able to show us, at a neurological level, what to look for in order

to identify high degrees of awareness, and thus, perhaps a Churchland-like theory of the future *could* tell the difference between Tarzan and Turing after all.

This paper began by outlining two questions, the answers to which are important to understanding (and accepting the plausibility of) Dennett's vNvm theory. The first question asked after a legitimate notion of explanation that would allow us, on the one hand, to grant that consciousness was (at times) serial and symbolic, and on the other, to remain committed to the idea that consciousness is supported by a brain which makes use of distributed representations and performs computations in parallel. I believe that sections (2) and (3) describe just such a notion: an endorsement of functionalism, where the isolation of 'what a system does' did not have to be reduced, or be even neatly mapped, onto an explanation of 'how the system works'. I did not have the space (or intention) to fully argue in favour of the framework I described, but hope nevertheless, to have demonstrated one defensible way of addressing the explanatory concerns of the first question. The secondary goal of these sections was to sketch Dennett's explanatory methodology such that the reader would understand the role of different aspects of his theory of consciousness. For example, in section (2.2 and 4.3) we interpreted 'von Neumannesque' as a description of a competency, pitched from the Intentional Stance (or at Marr level 1 if preferred), and in section (4.1 and 6.2) the Multiple Drafts model was seen as a Design Stance account; an answer to the 'how question' which remained neutral about underlying machinery. Once we uncovered Churchland's methodology in section (5.1) it was clear that his concept of explanation lacked an intermediary level (like the Design Stance or Marr level 2) between a description of a function or behaviour, and a mechanistic account of how this is realized. With this characterization, and the vantage point of our earlier sections (which allowed us to see Dennett's theory within the methodological framework underpinning it), some of Churchland's objections to the vNvm idea, turned out to be motivated more by his notion of what explanations should 'look like', than with errors or omissions within Dennett's theory itself.

The second question asked for ways in which a PDP device could develop a competency for von Neumannesque procedures. I could not be very conclusive in response to this

question, but I wanted to show that implementing Dennett's vNvm is possible, and further, possible in much the ways Dennett describes. In addressing this I began with a Marr-like approach: once we had isolated a fairly specific account of what happens (with the four features in section 4.3), I tried to show how each feature could conceivably be accounted for without having to find a physical serial processor, or physically discrete symbols. For each one, we considered both Dennett's ideas and some promising and complimentary theories from elsewhere. Every conclusion in this regard was tentative: I showed only how it *might* be the case that sequentiality is accomplished with relaxation networks (much like the Multiple Drafts model), how representational redescription *could* achieve the flexibility and manipulability of symbols, how this is *perhaps* how we develop the kind of informational organization necessary (under Dennett's view) for consciousness, and how interacting with the world, and internalisations of such interactions, constitutes a *plausible* controlling mechanism. Although all these are speculative conclusions, they nevertheless demonstrate that Dennett's conception of human consciousness as a serial simulation on a parallel architecture is one that is conceivably realizable.

Bibliography

Baars. B. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

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

Churchland, P. M. 1981. Eliminative Materialism and the Propositional Attitudes. In Lycan, W.G. ed. *Mind and Cognition*. Blackwell. 1990: 206-223.

Churchland P. M. 1988. *Matter and Consciousness*. MIT press.

Churchland, P. M. 1989. *A Neurocomputational Perspective*. MIT Press.

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

Churchland, P. M. 1999. Densmore and Dennett on Virtual Machines and Consciousness. *Philosophy and Phenomenological Research* 59(3): 762-767

Churchland, P. M. 2002. Catching Consciousness in a Recurrent Net. In Brook, A. and Ross, D. eds, *Daniel Dennett.* Cambridge University Press.

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

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

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

Clark, A. 1998. Magic Words: How Language Augments Human Computation. In Carruthers, P and Boucher, J Eds. *Language and Thought: Interdisciplinary Themes.* Cambridge University Press.

Clark, A., and Karmiloff-Smith, A. 1994. *The Cognizer's Innards: A Psychological and Philosophical Perspective on the Development of Thought:* Mind and Language 8: 487-519.

Crick, F. (1984) The Function of the Thalamic Reticular Complex: The Searchlight Hypothesis. *Proceedings of the National Academy of Sciences,* 81: 4586-4590.

Dawkins, R. 1976. *The Selfish Gene.* Oxford University Press.

Deacon, T. W. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain.* W.W. Norton and Company.

Dennett, D. 1978. *Brainstorms.* MIT Press.

Dennett, D. 1987. *The Intentional Stance.* MIT Press.

Dennett, D. 1991a. *Consciousness Explained.* Little Brown.

Dennett, D.1991b. Real Patterns. In *Brainchildren: Essays on Designing Minds.* MIT press. 1998: 95-120.

Dennett, D. 1992. Do-It-Yourself Understanding. In *Brainchildren: Essays on Designing Minds.* MIT press. 1998: 59-80.

Dennett, D. 1994a. Self-Portrait. In *Brainchildren: Essays on Designing Minds.* MIT press. 1998: 355-366.

Dennett, D. 1994b. Cognitive Science as Reverse Engineering: Several Meanings of "Top-Down" and "Bottom-Up". In *Brainchildren: Essays on Designing Minds*. MIT press. 1998: 249-261.

Dennett, D. 1994c. Labelling and Learning. *Mind and Language* 8: 54-48.

Dennett, D. 1995. Animal Consciousness: What Matters and Why. In *Brainchildren: Essays on Designing Minds*. MIT press. 1998: 337-350.

Dennett, D. 1998. Reflections on Language and Mind. In Carruthers, P and Boucher, J eds. *Language and Thought: Interdisciplinary Themes*. Cambridge University Press.

Densmore, S., and Dennett, D. 1999. The Virtues of Virtual Machines. *Philosophy and Phenomenological Research* 59(3): 747-61.

Edelman, G. M. 1992. *Bright Air, Brilliant Fire: On the Matter of the Mind.* Basic Books.

Elman, J. 1991. Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning* 7: 195-225.

Fodor, J. 1975. *The Language of Thought*. Crowell.

Fodor, J. 1981. *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* MIT Press.

Fodor, J., and Pylyshyn, Z. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. In Macdonald, C., and Macdonald, G. eds. *Connectionism: Debates on Psychological Explanation*. Blackwell. 1995: 90-163.

Jackson, F., and Pettit, P. 1988. Functionalism and Broad Content. *Mind* 97, (387): 381-400.

Karmiloff-Smith, A. 1985. Language and Cognitive Processes from a Developmental Perspective. *Language and Cognitive Processes* 1, no. 1: 61-85.

Karmiloff-Smith, A. 1986. From Metaprocess to Conscious Access: Evidence from Children's Metalinguistic and Repair Data. *Cognition* 23: 95-147.

Karmiloff-Smith, A. 1987. Beyond Modularity: A Developmental Perspective on Human Consciousness. Draft Manuscript of a talk given at the annual meeting of the British Psychological Society, Sussex, April.

Lloyd, D. 1989. *Simple Minds*. MIT Press.

Marr, D. 1982. *Vision.* W. H. Freeman and Co.

Miller, G. A. 1962. *Psychology: The Science of Mental Life*. Penguin Books Ltd.

Ross, D. 2000. Rainforest Realism: A Dennettian Theory of Existence. In Ross, D., Brook, A., and Tompson, D. eds. *Dennett's Philosophy: A Comprehensive Assessment.* MIT Press.

Ross, D., and Spurrett, D. (To appear). What to Say to a Sceptical Metaphysician: A Defence Manual for Cognitive and Behavioural Scientists. Forthcoming in: *Behavioural and Brain Sciences*.

Rumelhart, D., Smolensky, P., McClelland, J., and Hinton, G. (1986) Schemata and sequential thought processes in PDP models. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* volume 2, ed., J. McClelland et al. MIT Press.

Seager, W. 1999. *Theories of Consciousness.* Routledge.

Sejnowski, T., and Rosenberg, C. 1986. NETtalk: A Parallel Network that Learns to Read Aloud. John Hopkins University Technical Report JHU/EEC-86/01.

Selfridge, O. 1959. "Pandemonium: A Paradigm for Learning," *Symposium on the Mechanization of Thought Process.* HM Stationery Office.

Smolensky, P. 1988. On the Proper Treatment of Connectionism. In Macdonald, C., and Macdonald, G. eds. *Connectionism: Debates on Psychological Explanation.* Blackwell. 1995: 28-89.