From:

Michael Wheeler

Extended X: Recarving the Biological and Cognitive Joints of Nature

Draft Book Manuscript

NB: Please do not quote or cite without permission

Chapter 5 Making Your Mind Up

5.1 What Matters

Ask yourself the following question: why has the information processing account of cognition proven to be so popular with the fans of extended cognition? Here is one plausible answer. It's because information, in and of itself, is locationally uncommitted. Consider, for example, computational concepts, the sort of concepts in which any *worked out* information processing account of mind is most likely to be couched. Nothing in the basic notion of an algorithm privileges the inner. Thus imagine that a computational cognitive psychologist proposes an algorithmic solution to some cognitive task – say (to use our familiar example) performing long multiplication via pattern matching and symbol manipulation. *In principle* – that is, if we concentrate solely on the fundamental character of any algorithmic explanation – the various steps of that solution could be implemented inside the head, on paper using a pen, in a computer, or in a distributed system such that different steps are completed in each of these constituencies. Similarly, one might store and implement the rules of the algorithm in any of these regions. The algorithm *qua* algorithm simply doesn't care about its material instantiation.

This points us in the direction of a crucial observation. In order to get the idea of cognitive extension off the ground, the ExC theorist needs to embrace the *multiple realizability* of cognitive traits. To be more precise, the idea that cognitive traits are multiply realizable must make sense, if we are to argue for cognitive extension *by way of parity considerations*. Why? Because the parity principle is

based on the thought that it is possible for the very same type-identified cognitive trait to be available in non-extended and extended formats. Thus, in principle at least, the trait must be realizable in a purely organic medium or in one that involves an integrated combination of organic and non-organic structures. In other words, it must be multiply realizable. It is my view (to be defended in due course) that, given the current shape of things, the parity principle provides the *only* viable basis for ExC. If that is right then the multiple realizability of cognitive traits is a necessary component of ExC. We can put the key commitment here another way: when we specify a cognitive trait in a locationally uncommitted way, we implicitly endorse the idea that there is an equivalence class of different material structures and processes that might realize that trait. So ExC depends on the idea that the basic or first-order materiality of the systemic elements in question is relevant only insofar as it provides an explanation of how the equivalence classes in question may be implemented in the physical world. Let's call this property *implementational materiality*. ExC is committed to implementational materiality because it is committed to multiple realizability. And it is able to secure multiple realizability because it is committed to implementational materiality. In a sense, then, these two properties are opposite sides of the same conceptual coin.

Once the interlocked properties of multiple realizability and implementational materiality are brought to the fore, an eminently well-qualified candidate for a general approach to the nature of cognition that might preserve the ExC-related benefits of the information processing account, while avoiding the pitfalls of the specific versions of that account canvassed in the last chapter, presents itself. That candidate is *functionalism*. According to the traditional formulation of functionalism in the philosophy of mind, the canonical statement of which is arguably due to Putnam (1967), a mental state is constituted by the causal relations that the systemic state in question bears to sensory inputs, behavioural outputs, and other mental states. But depending on how one hears terms like 'sensory inputs' and 'behavioural outputs', this may betray a bias towards the inner that isn't, at root, a feature of functionalism's defining commitments. Fundamentally, the functionalist holds that what makes a systemic state a mental state is the set of causal relations that it bears to systemic inputs, systemic outputs, and other systemic states. Once we give this more general formulation of the functionalist line, we can allow the borders of the cognitive system to fall somewhere other than the sensory-motor interface of the organic body. And that opens the door to a cognitive system whose boundaries are located outside the skin. Thus we arrive straightforwardly at a position that Clark (2008a, b; see also Wheeler 2008) has dubbed *extended functionalism*. In the

rest of this chapter I shall endeavour to assemble a supporting case for this position.[1]

5.2 Putting the Function into Extended Functionalism

A little philosophical history will help to bring the extended functionalist landscape into better view. Functionalism (in its non-extended form) freed physicalist philosophy of mind from a kind of neural chauvinism. If our mental states were constituted by their functional roles, and the material contribution of our brains was merely implementational in character, then robots, Martians, Klingons, and gaseous creatures from the outer limits of the universe could all join us in having mental states, just so long as the physical stuff out of which they were made could implement the right functional profiles. Stretching the word 'skin' to include boundaries made of tin and gas, traditional functionalism bequeathed to the mind what we might call within-the-skin multiple realizability. And within-the-skin multiple realizability requires within-the-skin implementational materiality. But now extended functionalism merely plays out the same logic beyond the skin. If the specific materiality of the substrate doesn't matter to cognition, outside of the fact that it must be able to support the required functional profile, then what, in principle, is there to stop things-beyond-the-skin counting as proper parts of a cognitive architecture? Nothing, that's what. And this beyond-the-skin species of multiple realizability, which is just another way of characterizing the core philosophical commitment of ExC, requires beyond-the-skin implementational materiality. If we look at things this way, the really radical and revolutionary movement was functionalism, not ExC. ExC simply makes manifest one of the implications of functionalism. In other words, ExC is just a footnote to Putnam (Wheeler 2008).

Extended functionalism provides a general framework from which to approach anew the problem of supplying the ExC theorist with a theory-loaded, locationally uncommitted account of the cognitive. We have seen that the information processing approach promises the ExC theorist a route to multiple realizability and implementational materiality. Unfortunately, however, the raw idea that cognition is information processing is excessively liberal (it counts certain noncognitive phenomena as cognitive), while flagship attempts by ExC theorists to tighten up the notion so as to avoid such liberality either fail to achieve that aim or result in disproportionate elitism (they exclude genuinely cognitive traits from the domain of the cognitive). Can the functionalist

perspective help us to avoid these difficulties? Three considerations suggest that it can.

First, there is a clear sense in which, when it comes to the class of material systems that might count as cognitive, the baseline constraints introduced by the functionalist perspective impose more stringent restrictions than the baseline constraints introduced by the more general information processing approach. Only a subset of the systems that count as doing information processing, in the undemanding sense of that idea introduced earlier, will be organized into structured functional economies in which the functions and functional transitions within those economies are cognitively relevant.

Second, the most natural way to tighten up the information processing approach in an attempt to avoid excessive liberality is to add the constraint that the system in question should process information by building and transforming representations. (See e.g. Rowlands' arguments discussed above. Additional evidence comes from the representationalist assumptions that characterize the orthodox information processing approach in cognitive science.) However, as we have seen, this representation constraint is problematic, given that nonrepresentational processes such as CRC are likely to make functionally significant contributions within the mechanisms of cognition. If we strengthen the information processing approach by making the systematic manipulation of representations necessary for cognition, then we will be unable to create conceptual space for such nonrepresentational cognitive processes. The functionalist approach faces no such difficulty, since it harbours no in-principle opposition to the idea that, in some cases, cognitive functions may be underpinned by nonrepresentational vehicles.

As an example, consider the following functionalist gloss on an embodied-embedded model due to Randall Beer (2003). In this model, a simulated artificial agent controlled by a continuous time recurrent neural network is evolved to perform the minimally cognitive task of discriminating circular objects from diamond-shaped ones by catching the former but avoiding the latter. Beer argues that various standard aspects of the notion of representation, such as the semantic intelligibility of individual processing units within the behaviour-generating system, fail to characterize the underlying dynamical mechanisms designed by artificial evolution. Object-categorization is achieved through active processes of foveation and visual scanning, generated not by the regimented activity of individual inner processing units with identifiable semantically interpretations, but instead by shifting neuronal contributions that underpin a

global vectorfield of change in which sensory input is mapped to behavioural output in the context of current internal state. There is no conflict between this arguably nonrepresentational adaptive solution and a broadly functionalist perspective on the agent's cognitive architecture. The evolved agent is not some reactive or behaviourist stimulus-response machine. For one thing, which internal state the agent is already in, when it receives sensory input, differentially influences subsequent behaviour. In the most general terms, then, what it is to possess the minimally cognitive version of the capacity for object-categorization is to be capable of entering into certain internal states such that appropriate differential behaviours are produced in the presence of varying sensory inputs. That profile looks to be amenable to a broadly functionalist treatment.

The third consideration in favour of the functionalist perspective turns on the different senses that one might give to the term 'function'. As we learned during chapter 4, a further recommendation in Rowlands' recipe for tightening up the information processing account of what it is to be part of the cognitive involved binding the category of the cognitive to the property of having an adapted proper function. This move was open to counter-examples in the form of cognitive spandrels that have not been subject to secondary adaptation. It might seem that this worry about non-adapted traits will transfer to any functionalist theory of the cognitive. However, as long as functionalism isn't shackled to ultra-Darwinism by the assumption that *every* cognitive function must be understood in terms of some adapted proper function, the functionalist perspective will potentially be able to countenance the existence of non-adapted functions, and thus of non-adapted psychological phenomena. Of course, this requires that we have access to an appropriate alternative notion of function. The obvious candidate here is due to Robert Cummins (1975), who introduced a concept of function that, following Neander (1991; see also Amundson and Lander 1994), we can call *causal role function.* Causal role functions are identified not by evolutionary history, but via a strategy in which a scientist seeks to explain some interesting capacity of a system by showing how that capacity arises from the capacities of certain subsystemic elements and the interactions between them. Paul Griffiths (2005) illustrates the distinction in the domain of molecular biology. A "sequence of nucleotides GAU has the [adapted proper function] of coding for aspartic acid if that sequence evolved by natural selection because it had the effect of inserting that amino acid into some polypeptide in ancestral organisms" (ibid, p.1). The same nucleotide sequence "has the [causal role function] of coding for aspartic acid if that sequence has the

effect of inserting that amino acid into some polypeptide in the organism in which it occurs" (ibid, p.2).

At first sight it may look as if analyses of causal role functions will fail to distinguish between genuine functions and mere effects. This worry is invited by Griffiths' bald characterization of the causal role function of GAU. And indeed, critics of the causal role approach often submit alleged counter-examples that trade on this very idea. For example, Millikan (e.g. 1993, p.20) suggests that, according to the causal role account, clouds have the function of producing water in the rain cycle. Such consequences (which do not follow if we think in terms of adapted proper functions) would reflect badly on the causal role approach. Fortunately, the worry is (almost) wholly illusory, because Cummins (1975) gives us the conceptual resources to disregard, although not to prevent altogether, such analyses. He argues that a causal role functional analysis is ultimately valuable, and thus *worth* performing when, in the actual practice of science, a scientist takes some systemic capacity that she judges to be of interest and then (a) decomposes that capacity into an integrated suite of subsystemic capacities which are simpler and different in kind from it, and/or (b) demonstrates that the system in question is highly organized. Where the analysis exhibits such properties, our understanding of the target trait is deepened significantly. The idea that only valuable causal role functional analyses are worth performing allows us to disregard Millikan's alleged counter-example and others like it. For although it would be possible to attribute the causal role function of producing water in the rain cycle to a cloud, that analysis would fail to meet either of Cummins' value conditions, and so would be scientifically worthless.

By contrast, here are some illustrative examples of valuable causal role functional analyses, the first from Amundson and Lauder (1994). A functional anatomist might provide a valuable causal role functional analysis of the crushing capacity of the jaw of a particular creature by, in part, describing the capacity of a particular muscle to contract and thus to bring closer together two bones of the jaw. If that capacity were combined with other subsystemic capacities, to form an integrated account of how the jaw crushes things, then it would be appropriate by Cummins' criteria to attribute to the muscle in question the causal role function of bringing the two bones closer together. Similarly, a Marrian (Marr 1982) vision psychologist might provide a valuable causal role functional analysis of the ability of human beings to see the 3-D shapes of the objects in the visual field by, in part, describing the capacity of a subsystem to take as its input a 2-D array of light intensity values recorded at

the retina, and to compute a 2-D representation of the lines, closed curves, and termination-points of discontinuities in that array. When that capacity is combined with other subsystemic capacities, to form an integrated account of how the visual system arrives at a representation of the 3-D shapes of the objects in the visual field, then it would be appropriate by Cummins' criteria to attribute to the subsystem in question the causal role function of computing a 2-D representation of the structure of the discontinuities in the retinal array. Finally, at a more abstract level, consider the generic practice of homuncular analysis (see chapter 4), in which a system is analyzed into a set of hierarchically organized, communicating modules, each of which performs a well-defined sub-task that contributes towards the collective achievement of a global systemic solution. Unless one explicitly introduces Darwinian selection into this picture in order to specify evolutionary purposes for the various systems and subsystems (a move which is surely not present in the vast majority of such analyses, and which would involve a problematic attribution of adapted proper functions at the subsystemic level – see next paragraph), homuncular analysis is most naturally interpreted as a species of causal role functional analysis.

A full examination of the debate over the status of causal role functions in biology is beyond the scope of the present treatment (for more details, see e.g. Neander 1991, Millikan 1993, Sober 1993, Amundson and Lauder 1994, Sterelny and Griffiths 1999). It is time to apply what we have learned. It may well be true that, in the real world, most of the capacities for which scientists produce valuable causal role functional analyses are capacities that have undergone a process of design by natural selection, and thus have adapted proper functions. Notice, however, that it doesn't follow from this that the subsystemic capacities that figure in such analyses, the capacities to which the analyses attribute causal role functions, have adapted proper functions in addition to causal role functions. As Amundson and Lander, again, rightly observe, "the generalization *Functionally complex items have selective histories* does not by itself imply that a positive selective influence was responsible for every causal property of every component of the functional complex" (ibid. p.459). What this tells us is that there is no one-to-one correspondence between adapted proper functions and causal role functions, a result which opens the door to the following further possibility: there may be valuable causal role functional analyses of non-selected-for traits, traits such as spandrels that have not been subject to secondary selection.

To make this possibility vivid, consider the fact that genes are sometimes linked physically, in such a way that the evolutionary fate of one gene is bound up

with the evolutionary fate of another. This provides the basis for a phenomenon known as *genetic hitchhiking*. Here is a fictitious hitchhiking example that I have used before (Wheeler 2003, 2005b, 2006) and that I shall use again later in this book (chapter 7). Assume that, in some creature, the gene for a thick coat is linked to the gene for blue eyes. Let's also assume that this creature lives in an environment in which it is selectively advantageous to have a thick coat, and selectively neutral to have blue eyes. What will happen is that the gene for a thick coat will be selected for. But since the gene for blue eyes is linked physically to the gene for a thick coat, the gene for blue eyes will be inherited too, even though it bestows no selective advantage, has not been selected for, and thus has no adapted proper function. The trait of having blue eyes is thus a spandrel, a selectively neutral by-product of having a selectively advantageous thick coat. By hypothesis, this spandrel has not been subject to secondary selection. Nevertheless, there seems to be little doubt that a suitably trained biologist could provide a valuable causal role functional analysis of the developmental mechanisms by which that trait is generated in each individual bearer. In part, this analysis would describe the capacity of the causally relevant gene to produce certain developmentally downstream effects during protein synthesis, effects that ultimately result in blue eyes being installed in the phenotype.

At this point a critic might complain that even if there is a valuable causal role functional analysis for at least some selectively neutral traits, there will be no such analysis available for selectively disadvantageous traits, such as the psychological tendency for misattribution in human memory. But this critic has simply failed to purge himself of the adaptationist inclination to interpret function in terms of purpose. Once we resist that inclination, there is no reason to think that the tendency for misattribution is anything other than a scientifically interesting capacity of human beings that will succumb to a valuable causal role functional analysis. For example, a particular form of misattribution, known as false recognition, occurs when individuals mistakenly claim that a novel item or event is familiar. On the basis of work by McClelland (1995; discussed by Schacter and Dodson 2001), one might plausibly suggest that a valuable causal role functional analysis of the capacity for false recognition would, in part, describe the capacity of a psychological subsystem to extract the gist of experiences, a capacity that, because of its intrinsic tendency towards distortion, results in the capacity for false recognition, when combined with certain storage and retrieval subsystems. Of course, when combined with other kinds of subsystem, the very same gist-extracting subsystem might be mentioned in a valuable causal role functional analysis of a capacity such as

generalization across tasks, a capacity that might itself possess an adapted proper function.

5.3 Sleeping with the Enemy (Part Two): What Cognitive Scientists Care About

With the 'function' in 'extended functionalism' now understood fundamentally in terms of causal role function, how far have we come in our attempt to sketch the benefits (for the ExC theorist) of a functionalist framework? The answer is: not far enough. I claimed earlier that extended functionalism avoids the excessive liberality problem that blighted the raw information processing approach, precisely because only a subset of the systems that count as doing information processing will be organized into structured functional economies in which the functions and functional transitions within those economies are cognitively relevant. However, as sharp-eyed readers might have noticed, there is a sense in which this claim assumes that the extended functionalist has some way of determining which of the many functions performed by various complexes of organic machinery coupled with (in some cases) non-organic machinery are the cognitively relevant ones. In other words, from what we have seen so far, there is a sense in which extended functionalism *assumes* an account of the cognitive; it does not itself *supply* one.

The root issue here becomes visible once one takes seriously the commitments that extended functionalism inherits from Cummins-style causal role functional analysis. In the course of our development of extended functionalism as a position, we have, in effect, left behind the old philosophical idea of functionalism as a kind of conceptual analysis of ordinary language mental vocabulary. Instead, the concept of causal role functional analysis offers us a philosophical articulation of an explanatory strategy adopted by certain scientists. And the first step of that strategy (see above) is to fix on some systemic capacity of interest within the domain of science in question (e.g. the domain of functional anatomy, developmental biology or cognitive science). So, the scientifically illuminating decomposition of a target capacity into a suite of highly organized, interacting, simpler subsystems possessing causal role functions takes place *after* the question of what counts as part of the domain of explanatory interest has already been settled. Thus the causal role functional analysis of the developmental system that underlies the blue-eyes trait in our imaginary creature depends on the prior identification of that trait as a biological developmental phenomenon. And the causal role functional analysis

of the cognitive system that underlies the capacity of human beings to see a world of 3-D objects depends on the prior identification of that capacity as a cognitive phenomenon.

We are now on the verge of being reconnected with a position in the ExC debate that we thought we had put behind us. To see why, notice that there is nothing about the process of prior identification just highlighted that requires us to step outside the explanatory practices of the sciences in question. As I have argued elsewhere (Wheeler 2005b, chapter 5), in *any* scientific investigation – and that includes, of course, the scientific investigation of cognition – the scientist will make certain, often implicit assumptions about the constitutive character of the target phenomena, assumptions that are in principle sensitive to, and thus modifiable in the wake of, the success or failure of the empirical models that those assumptions support. There is a job for philosophy to do in articulating, amplifying and examining the coherence of those assumptions, but nevertheless they are there already, albeit in a partially buried form, in the conceptual foundations that shape scientific theorizing and research. Such assumptions plausibly form the basis of the decisions that scientists routinely make in selecting phenomena for subsequent causal role functional analysis. This suggests that if we want to isolate a scientifically informed, theory-loaded, locationally uncommitted account of the domain of the cognitive, we could do worse than to allow that domain to be delineated by the range of capacities that the relevant scientists care about, plus (derivatively) the causal role functions performed by the subsystemic elements whose organization and interaction underlie those capacities. But now a problem looms. The foregoing picture might be taken to suggest that the domain of the cognitive should be fixed by reference to the capacities, whatever they may be, that cognitive psychologists care about. And that, of course, was one consideration used by Rupert, Adams and Aizawa to argue that cognition remains a resolutely skin-side phenomenon (see chapter 3). Once again, then, we seem to find ourselves rather too close for comfort to a position occupied by the friendly hostiles.

Fortunately, there is an important corrective to the analysis on offer from Rupert, Adams and Aizawa that, once implemented within our emerging ExC framework, restores a healthy distance between the two perspectives. For it is arguable that these critics have a skewed view of where the intellectual core of cognitive science is located in disciplinary space. To explain: as I have pointed out already, the critics take it that cognitive science is represented principally, and sometimes it seems exhaustively, by orthodox, inner-oriented cognitive psychology. It is little wonder, then, that, in their hands, an appeal to the

methods and practices of cognitive science seems to tell against the possibility of cognitive extension. But there is another way of thinking about the character of cognitive science, one justified by both the history and the philosophical character of the field, that, as we shall see, promises to be rather more ExC-friendly. According to this alternative view, the intellectual core of cognitive science is not orthodox cognitive psychology, but rather *artificial intelligence (AI)* in all its various forms (cybernetic, classical, connectionist, embodied, situated). (For expressions of this idea, see Boden 1990, p.1; Wheeler 2005b, p.1. See also Boden 2006 chapter 4 on the importance of McCulloch's and Pitts' seminal 1943 paper, 'A Logical Calculus of the Ideas Immanent in Nervous Activity'. This paper showed that it was possible for certain kinds of neural network to compute certain specified logical functions. Boden judges it to be the "common birthplace of traditional and connectionist AI" (Boden 1990, p.2) and the "first paper in cognitive science" (Boden 2006, p.702).)

In my view, there are two main independent (of ExC) reasons for adopting this AI-centric vision of cognitive science. First, AI has been the source of many of cognitive science's (and thus cognitive psychology's) most cherished concepts and models. This explanatory toolkit includes theoretical notions, such as that of an algorithm, as well as classes of psychological models, such as parallel distributed processing networks and, more recently, subsumption architectures (Brooks 1991; see Prescott et al. 1999 for evidence of subsumption-like structures in animal psychology). Secondly, as Boden (1990, p.1) observes, it is possible to conceive of AI as the *science of intelligence in general*. Boden explicates this idea in the following way:

> As such, [the] goal [of AI] is to provide a systematic theory that can explain (and perhaps enable us to replicate) both the general categories of intentionality and the diverse psychological capacities grounded in them. It must encompass not only the psychology of terrestrial creatures, but the entire range of possible minds. It must tell us whether intelligence can be embodied only in systems whose basic architecture is brainlike (involving parallel-processing within networks of associated cells), or whether it can be implemented in some other manner. (Boden 1990, p.1)

On the basis of this analysis, one might argue that it is by placing AI at the centre of cognitive science that we are able to distinguish the latter enterprise from its more narrowly focussed component discipline of cognitive psychology. For if AI is the intellectual heart of cognitive science, then Boden's analysis of AI

may be generalized, allowing cognitive science to be depicted as a distinctive intellectual endeavour whose goal is a *science of cognition in general*. The scope of such a science will include extant examples of terrestrial cognition (including human cognition), possible forms of terrestrial cognition, and possible forms of alien cognition – in short, the "entire range of possible minds". Indeed, as long as one doesn't understand the notion of a cell in an overly biochemical way, even Boden's final observation that cognition might, in principle, be restricted to brainlike architectures (brain-*like* architectures, one should note, not brains), invites an abstract specification of those architectures, one given in terms of parallel processing networks that might themselves instantiate a range of processing solutions not presently found in human psychology. Once the goal of explaining all possible forms of cognition is revealed as the distinguishing mark of cognitive science, the resulting intellectual animal seems to have little in common with the chauvinistic beast unleashed into the ExC debate by the friendly hostiles. Indeed, a science of cognition in general would seem to be a likely source for precisely the kind of scientifically informed, theory-loaded, locationally independent account of the cognitive that the ExC theorist needs. In the next section I shall try to give some substance to this idea.[2]

5.4 Extended Physical Symbol Systems

Newell and Simon, two of the early architects of AI, famously claimed that a "physical symbol system has the necessary and sufficient means for general intelligent action" (Newell and Simon 1976, p.116). As anyone familiar with cognitive science will know, a physical symbol system is (roughly) a classical computational system instantiated in the physical world, where a classical computational system is (roughly) a system in which atomic symbols are combined and manipulated by structure sensitive processes in accordance with a language-like combinatorial syntax and semantics. I shall take it that the phrase 'means for general intelligent action' points to a kind of cognitive processing. More specifically it signals the sort of cognitive processing that underlies "the same scope of intelligence as we see in human action… in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity" (Newell and Simon 1976, p.116). What we are concerned with, then, is a *human-scope cognitive system*. Notice that the concept of a human-*scope* cognitive system is not a species-chauvinistic notion. What matters is that the system exhibit roughly the same degree of adaptive flexibility we see in humans, not that it have our particular biological make-up, species ancestry or

developmental enculturation. Newell and Simon's physical symbol systems hypothesis may therefore be unpacked as the dual claims that (a) any human-scope cognitive system will be a physical symbol system, and (b) any physical symbol system of sufficient complexity may be organized so as to be a human-scope cognitive system.

Understood this way, the physical symbol systems hypothesis advances a scientifically informed, theory-loaded account of the (human-scope) cognitive, one that supports a computational form of functionalist theorizing. But can it tick all our boxes by being a locationally independent account too? The answer, it seems, is yes. For while classical cognitive scientists in general thought of the symbol systems in question as being realized inside the head, there is nothing in the basic concept of a physical symbol system that rules out the possibility of extended material implementations. In fact, as we are about to see, the notion of an extended physical symbol system is one way of expanding on an idea that we have encountered already.

Recall once again our stock example of how ordinary human competence in the task of long multiplication may be the outcome of a canny partnership between inner neural processes, sensorimotor skills, and external, materially realized symbolic structures. The central insight of this example is generalized and deepened by William Bechtel, in a series of compelling treatments that combine philosophical reflection with empirical modelling studies (Bechtel 1994, 1996; Bechtel and Abrahamsen 1991). Building on the connectionist orientation of Rumelhart et al.'s (1986) treatment of the long multiplication case, Bechtel develops and defends the view that certain human-scope cognitive achievements, such as mathematical reasoning, natural language processing and natural deduction, are the result of sensorimotor-mediated interactions between internal connectionist networks and external symbol systems, where the latter feature various forms of combinatorial syntax and semantics.

It is useful to approach Bechtel's suggestion (as he does himself) by way of Fodor and Pylyshyn's (1988) well-known claim that connectionist theorizing about the mind is, at best, no more than a good explanation of how classical states and processes may be implemented in neural systems. Here is a brief reminder of Fodor and Pylyshyn's key argument. It begins with the empirical observation that thought is *systematic*. In other words, the ability to have some thoughts (e.g. that Elsie loves Murray) is intrinsically connected to the ability to have certain other thoughts (e.g. that Murray loves Elsie). If we have a classical vision of mind, the systematicity of thought is straightforwardly explained by

the combinatorial syntax and semantics of the cognitive representational system. The intrinsic connectedness of the different thoughts in question results from the fact that the processing architecture contains a set of atomic symbols alongside certain syntactic rules for recombining those symbols into different molecular expressions. Now, Fodor and Pylyshyn argue that although there is a sense in which connectionist networks instantiate structured states (e.g. distributed connectionist representations have active units as parts), *combinatorial* structure is not an essential or a fundamental property of those states. This leaves connectionist networks inherently incapable of explaining the systematicity of thought, and thus of explaining thinking. What such systems might do, however, is explain how a classical computational architecture may be implemented in an organic brain.

Bechtel agrees with Fodor and Pylyshyn on two key points: first, that where systematicity is present, it is to be explained by combinatorially structured representations; and secondly, that connectionist networks fail to realize combinatorial structure. He does not need to endorse Fodor and Pylyshyn's claim that all thought is systematic, however. For his purposes, all that is required is that some cognitive activities (e.g. linguistic behaviour, natural deduction, mathematical reasoning) exhibit systematicity. Against this background, Bechtel's distinctive move is to locate the necessary combinatorial structure in systems of representations that remain *external* to the connectionist network itself. Given the idea that our inner psychology should be conceived in connectionist terms, this is tantamount to saying that the necessary combinatorial structure resides not in our internal processing engine, but rather in public systems of external representations (e.g. written or spoken language, mathematical notations). As Bechtel (1994, p.4, manuscript) himself puts it, the "property of systematicity, and the compositional syntax and semantics that underlie that property, might best be attributed to natural languages themselves but not to the mental mechanisms involved in language use".

For this idea to fly, it must be possible for the natural sensitivity to statistical patterns that we find in orthodox connectionist networks to be deployed in such a way that some of those networks, when in interaction with specific external symbol systems, may come to respect the constraints of a compositional syntax, even though their own inner representations are not so structured. Bechtel's studies suggest that this may be achieved by exploiting factors such as the capacity of connectionist networks to recognize and generalize from patterns in bodies of training data (e.g. large numbers of correct derivations in sentential arguments), plus the temporal constraints that characterize real embodied

engagements with stretches of external symbol structures (e.g. different parts of the input will be available to the network at different times, due to the restrictions imposed by temporal processing windows). The conclusion is that "by dividing the labor between external symbols which must conform to syntactical principles and a cognitive system which is sensitive to those constraints without itself employing syntactically structured representations, one can perhaps explain the systematicity… of cognitive performance" (Bechtel 1994, p.6, manuscript).

How should we interpret the distributed solutions that Bechtel favours – as examples of embodied-embedded cognition or as instances of cognitive extension? Bechtel himself stops short of the extended option. Thus, as we have just seen, he tellingly describes systematicity as a feature of "cognitive *performance*" rather than as a property of the cognitive system, and states that the compositional syntax and semantics "might best be attributed to natural languages themselves *but not to the mental mechanisms involved in language use*" (my emphases). What this indicates is that, for Bechtel, the genuinely cognitive part of the proposed solution remains skin-side. Let's see what interpretation we get, however, once we apply the parity principle. If the envisaged system of syntax-sensitive processes and combinatorially structured symbols were all stuffed inside the agent's head, we would, I think, have no hesitation in judging the symbol structures themselves to be bona fide parts of the agent's cognitive architecture. Equality of treatment therefore seems to demand that the external symbol structures that figure in the functionally equivalent distributed version of that solution also be granted cognitive status. On the strength of the parity principle, then, what we have here are models of extended cognition. (In previous treatments of Bechtel's logical reasoning studies from an ExC perspective, Rowlands (1999, pp.168-171) and Menary (2007, also pp.168-171) rely at root not on parity considerations to justify the claim of cognitive extension, but rather on the integration of inner connectionist processing with external symbol systems in order to complete a cognitive task that could not ordinarily be achieved by the inner networks alone. I have already suggested that this sort of view falls prey to an excessive liberality worry, but I shall return to the issues again later.)

In truth, the foregoing direct appeal to parity considerations takes us only part of the way to the distinctive ExC conclusion. As we have seen, parity-based arguments remain inconclusive until they receive backing from some theory-loaded, locationally uncommitted account of the cognitive that sets the benchmark for parity. And that's where our AI-centred cognitive science – our

science of cognition in general – is supposed to come in. To illustrate how the physical symbol systems hypothesis, in particular, might provide the basis for such a contribution, we need to take two steps. First we need to weaken that hypothesis so that its scope is limited to *subsystems* of an overall cognitive architecture. This respects the idea, already expressed in our coverage of Bechtel's work, that only *some* cognitive achievements will be explained by combinatorial symbol manipulation. As we might now put it, a physical symbol system, when placed in the operating context of a complete cognitive architecture, has the necessary and sufficient means for *certain aspects of* human-scope intelligent action. The second step is to claim that both the wholly inner and the environment-involving versions of the Bechtel-style network-plus-symbol-system architecture are instantiations of sufficiently complex and suitably organized physical symbol systems (as reconceived). The best way to appreciate the plausibility of this claim is to consider the most obvious objection against it.

Many cognitive scientists faced with the present suggestion will want to complain that the kinds of pattern-matching and pattern-completion processes realized by connectionist networks are not equivalent to the syntactic rules present in classical architectures. With all due respect this is, I think, a failure of the imagination. It is of course true that the network processes concerned are not *explicitly* rule-driven in a classical sense, but two considerations strongly indicate that this is not the end of the matter. First, the keystone of Bechtel's model is the thought that the networks involved are genuinely sensitive to the constraints of a compositional syntax. Thus, pending good arguments to the contrary, one might insist that Bechtel's networks *implicitly* realize the rules in question, at least in the minimal sense that, in this case (although not in others), classical-style rules will provide a perfectly reasonable, high-level, idealized description of the network's processing activity. (The fact that there is idealization here should not concern anyone. For one thing, idealization is part of scientific explanation. For another, orthodox connectionist models are themselves abstract idealizations of real brains.) Secondly, and from a more radical perspective, it may be that the classical rules are not implicitly realized in the neural network alone. If we think of those rules as principles that, at an agential level, govern the skilled embodied manipulations of certain external material symbols, it might be more accurate to think in terms of dynamic subagential vehicles that include not just neurally-implemented connectionist elements, but also non-neural bodily factors, including physical movements.

At this juncture it will be useful to consider Bechtel's own response to the suggestion that his models constitute implicit realizations of classical architectures. He writes:

> [It] would be an exercise in futility to try to prove that all of the explicit activity generated by one of my networks in solving a problem corresponds exactly to a series of implicit symbolic operations. It is neither necessary nor desirable that this be the case. Instead, I would point again to the advantages of a division of labor between language-like external representations and neural-like internal processing. To do its job, the internal processing system need not replicate the structure of an external representation that is made available on its input units; its knowledge of language and logic is procedural, cooperative, and sensitive to composition without being compositional. We would all like to attain a better understanding of the internal operations of networks, but focusing our search on functional equivalents to symbolic operations could keep us from noticing what is most worth seeing. (Bechtel 1994, p.23, manuscript)

What is striking about Bechtel's response is the way in which it assumes that, for there to be a classical interpretation of his models, it would somehow need to be correct to say that, for each example, an entire system of rules plus symbols is implicitly realized *within the network itself*. If that were true, then Bechtel would, I think, be right that any such interpretation is "neither necessary nor desirable". The networks in question are not by themselves implicit physical symbol systems. However, my way of thinking about these models does not make the key internalistic assumption. "[F]unctional equivalents to symbolic operations" will be found not in the internal activity of the networks, but in the activity of distributed systems, each of which encompasses a network, a non-neural body, and an external symbol system. There is no requirement that the networks implicitly realize compositionally structured representations, only that they implicitly realize structure-sensitive rules. If this is the right way to think, then what we have here are models of extended physical symbol systems, and thus of extended cognition.

5.5 A Parting of Ways

Before anyone gets carried away, I should stress the following points. It may be that, ultimately, there are decisive objections against the idea of connectionist

networks (or connectionist networks plus embodied manipulations) implicitly realizing classical-style structure-sensitive rules. Alternatively it may be that, in the end, nothing in cognition, extended or otherwise, answers to the description of a physical symbol system. No matter. The foregoing example was intended purely as an *in-principle illustration* of the fact that an AI-centred cognitive science is well-placed to provide us with a theory-loaded, locationally independent account of the cognitive. That point does not stand or fall with the details of the specific illustrative example.

One might think that this attempt to distance myself from the details of the physical symbol systems hypothesis is more than a smidgen too quick, and that the crucial property of locational independence is parasitic on a conception of cognition that classical theories support and that other approaches in AI don't. Thus it might seem that it is the classical commitment to an essentially language-like representational medium that opens the door to the presence of non-organic elements in the vehicles of cognition, by establishing an immunity to body-centred structural constraints. This is an immunity which, so the objection goes, is not sustained by certain other approaches AI. Here one might mention connectionism with its much-vaunted neural inspiration, or situated robotics with its stress on the ways in which corporeal design or bodily movement may change the character of certain adaptive problems. (See Harvey et al. 1994 and Webb 1994 for classic examples of how situated robotics makes use of corporeal design. The first involves the spatial arrangement of visual receptors, the second the propagation of sound through a tracheal tube. Webb's robot is described in chapter 1 above. A robotic example of how bodily movement may restructure a cognitive problem space is described below.) The suspicion, then, is that something about the way in which these prominent conceptual frameworks for AI research appeal directly to the agent's embodiment means that any resulting theory-loaded account of the cognitive will be locationally committed in a (skin-side) way that impedes cognitive extension.

In considering this worry let's put aside, for the moment, the seemingly undeniable fact that there are important dimensions along which most connectionist networks and their real neural relations diverge radically, a fact which would tend to destabilize the claim that connectionist theory appeals to the details of organic embodiment in such a way that cognitive extension is straightforwardly ruled out.[3] For there is a deeper issue vying for our attention here, one that places the distinction between embodied cognition and cognitive extension in a new, and altogether more dramatic, light. That issue is how, in

general terms, we conceptualize the fundamental contribution of the body to cognitive phenomena. The two strands of thinking are nicely untangled by Andy Clark.

> One of those strands depicts the body as intrinsically special, and the details of a creature's embodiment as a major and abiding constraint on the nature of its mind: a kind of new-wave body-centrism. The other depicts the body as just one element in a kind of equal-partners dance between brain, body and world, with the nature of the mind fixed by the overall balance thus achieved: a kind of extended functionalism (now with an even broader canvas for multiple realizability than ever before). (Clark 2008, pp.56-57; for a complementary analysis, see Wheeler 2008.)

To the extent that the body is conceptualized as 'no more than' a bridge to new functional organizations, its materiality (here interpreted as the fine-grained details of its wiring and biochemistry) is implementational in nature, and the resulting theoretical position guarantees the degree of multiple realizability required for ExC to be true. To the extent that the body is conceptualized as making a special, non-substitutable contribution to cognition (generating what Clark 2008, p.50, calls "total implementation sensitivity"), its materiality is, as I have called it elsewhere (Wheeler 2008) *vital*, and any degree of multiple realizability is, if not ruled out altogether, at least severely curtailed (see e.g. Shapiro 2004, especially p.167). The latter vision of how embodiment contributes to cognition would be justified if, for example, having the very same experience required having the very same fine-grained bodily structure (as sometimes argued by Noë, e.g. 1994, p.112). It seems to me, however, that the vision that pervades just about every corner of AI research is of the former, and not the latter, kind. Here is some evidence.

It is often noted that the cognitively relevant functions implemented by connectionist networks may often need to be specified at a finer level of grain than those performed by classical computational systems. Thus, in the case of connectionist networks, the salient functional roles will often be specified in terms of mathematical relations (between units) that do not respect the boundaries of linguistic or conceptual thought. However, this state of affairs gives us no reason to think that the distinctive contributions of those networks are not fully exhausted by those functional roles. That's why Clark (1989, 1999) has described connectionist theory as a kind of *microfunctionalism*. Significantly, microfunctionalist connectionism "specifies a system only in terms of input-

output profiles for individual units and thus is not crucially dependent on any particular biological substrate" (Clark 1999, p.40). This clears the way to alternative realizations of the functions in question, including not merely different biological (organic) instantiations, as Clark's text directly suggests, but also extended implementations.

What about the appeal made to bodily factors (such as spatial organization and physical movement) in situated robotics? Does that undermine the claim that AI might provide a conceptual basis for cognitive extension? Here is an example which suggests that the answer is 'no'. Clark and Thornton (1997) claim that there are certain learning problems – so-called type-2 problems – where the target regularities are inherently relational in nature, and so are statistically invisible in the raw input data. In the face of such problems, Clark and Thornton urge the use of general inner processing strategies that systematically re-represent the input data, such that representation is traded against complex computational search to render the learning problem tractable. As Scheier and Pfeifer (1998) demonstrate, however, type-2 problems may be solved by a process in which a mobile agent uses autonomous bodily motion to actively structure input from the environment. In this way, an intractable type-2 problem may be transformed into a tractable type-1 problem (one in which the target regularity is non-relational and thus visible in the raw input data). For example, the type-2 problem presented by the task of avoiding small cylinders while staying close to large ones was overcome by some relatively simple evolved neural network robot controllers. Analysis demonstrated that most of these controllers had evolved a systematic circling behaviour which, by inducing cyclic regularities into the input data, turned a hostile type-2 climb into a type-1 walk in the park. In other words, adaptive success in a type-2 scenario (as initially encountered) was secured not by inner re-representation, but by an approach in which the agent, "by exploiting its body and through the interaction with the environment ... can actually generate ... correlated data that has the property that it can be easily learned" (Scheier and Pfeifer 1998, p.32).

The key point, for present purposes, is that the re-structuring of the learning problem achieved by the bodily movements of Scheier and Pfeifer's robot is functionally equivalent to the transformation of that problem effected by Clark and Thornton's inner re-representation strategy. Thus it makes sense to think in terms of a multiply realizable functional role, one that may be performed by inner mechanisms, by bodily movements, or, taking things one step further, by certain extended systems. For example, it is surely conceivable that the systematic recoding of type-2 task data, so as to generate a tractable type-1

learning problem, may be achieved using an external device such as suitably programmed mobile computing device. If any other parity considerations that are relevant here are also met, that device may become an integrated part of an ecological vehicle of cognition. Whatever account of the cognitive is at work in situated robotics, it looks to be locationally uncommitted.

I have argued that an AI-centred cognitive science, interpreted as a science of cognition in general, is an appropriate and potentially powerful place to look for a scientifically informed, theory-loaded, locationally uncommitted account of the cognitive (the benchmark for parity that ExC needs). However, in developing the case for this conclusion, I have presided over a final parting of the ways between, on the one hand, ExC (understood as involving an extended functionalist commitment to a kind of open-ended multiple realizability) and, on the other, a style of embodied-embedded theorizing about the mind that depicts the organic body as, in some way, intrinsically special in the generation of cognitive phenomena. At root this divorce within current thinking is driven by a fundamental disagreement over how philosophy and cognitive science should conceive of the materiality of the body, as just one implementating substrate among possible others, or as a vital and irreplaceable determinant of cognitive life. This is a disagreement that, later in our investigation, will demand our attention again.

5.6 Troubles for Extended Functionalism?

Having introduced the idea that ExC ought to be developed within a broadly functionalist framework, it is time to consider three arguments (two in this section and one in the next) that critically engage, in one way or another, with the relationship between functionalism and cognitive extension.

Adams and Aizawa (2008, pp.68-70) argue that we should expect the vehicles of cognition to be exclusively neuronal in character, because we should expect processes as distinctive as cognitive processes to be realized by correspondingly distinctive lower-level processes. The latter expectation is in turn to be justified by the general principle that "[r]oughly speaking, lower-level processes should be as distinctive as the higher-level processes they realize" (ibid., p.68). As evidence for the way in which this principle plausibly isolates neuronal states and processes as the only vehicles of cognition, Adams and Aizawa point to the manifest differences between two sets of lower-level vision-related processes that are instantiated on either side of a transduction interface positioned at the

retina. Thus in the eye, prior to the retina (e.g. in the cornea and the lens), we find optical processes essentially similar to those present in non-organic optical machinery. When light enters the retina, however, there is a shift to molecular processes that, among other things, result in the colour-sensitive, orientation-sensitive and motion-sensitive selective release of neurotransmitters. According to Adams and Aizawa, this transition in lower-level processes also marks a transition from the noncognitive to the cognitive.

Adams and Aizawa's argument implies a rejection of the claim that human cognitive traits are multiply realized. It is here that a critical engagement with functionalism ensues. Adams and Aizawa write: "Functionalists about cognition might… observe that, in principle, anything could be organized in such a way as to give rise to cognitive processing. But our point is that, even though many things *could*, in principle, be organized to form a cognitive processor, it is reasonable to conjecture that only neuronal processes are in fact so organized" (ibid., p.69). This tells us something useful. Understood as part of a general critique of cognitive extension, now interpreted in terms of extended functionalism, Adams and Aizawa's argument is levelled not against the modal versions of ExC, but against the non-modal version (see chapter 2 for formulations of these different versions). In view of this, one defensive strategy that the ExC theorist might pursue would be to present scientific evidence for the following claim: there are extant biological examples which falsify the background principle that lower-level processes should be as distinctive as the higher-level processes they realize. If this background principle is false, we have no general reason to expect each distinctive higher-level biological process to be implemented in a single material substrate, and thus no general reason to expect cognition to be realized exclusively in a neuronal substrate. What the ExC theorist needs to do, then, is find scientific evidence of a distinctive biological process that is, *in fact*, multiply realized by more than one kind of lower-level process.

As it happens, it seems that such evidence is plentiful, in examples of what is known as *functional convergence in evolution*. Convergent evolution is a widespread phenomenon in which a particular biological trait evolves independently in more than one lineage, from different ancestors. One kind of convergent evolution involves functional convergence (Doolittle 1994), a process in which two or more biological entities perform the same function, but do so by way of entirely different underlying structures and mechanisms. Here is an example of functional convergence in molecular evolution. Alcohol dehydrogenases are enzymes that, in humans and many other animals, break

down alcohols that might otherwise be dangerous. They figure in the molecular economies of vertebrates and fruit-flies, and perform functionally equivalent roles in each of these biological contexts, but the vertebrate enzymes and the fruit-fly enzymes display no sequence similarity with each other, have fundamentally different tertiary structures, and catalyze alcohol into acetaldehyde using different chemical reactions (Doolittle 1994). This is just one example of the fact that there exist distinctive biological phenomena that are multiply realized. Adams and Aizawa's background principle is false. Extended functionalist minds may yet be actual.

A second argument that, in a different way, targets extended functionalism hails from Mark Rowlands (manuscript). This argument emerges out of a consideration of Rupert's (2004) claim that the processes involved in putative cases of extended memory differ in such fundamental ways from those involved in cases of ordinary internal memory that the extended cases cannot count as cognitive. In chapter 3, I suggested that Rupert's claim depends on the assumption that what counts as cognition should be fixed by the fine-grained profile of the inner, an assumption that begs the question against ExC (Wheeler 2008). However, as I also noted then, Rowlands has subsequently pointed out that if all the ExC theorist does in replying to Rupert is make the counter-claim that what counts as cognition should *not* be fixed by the fine-grained profile of the inner, then the question-begging simply travels in the other direction. The result is a deadlock. In fact, however, Rowlands' analysis is somewhat more nuanced than my first pass over the issue indicated, because he also suggests that the deadlock in question is, in truth, a disguised version of a familiar impasse in functionalist philosophy of mind. Here is the relevant passage from Rowlands:

> If Rupert's arguments against the extended mind are question-begging because they presuppose a chauvinistic form of functionalism, it is difficult to see why arguments for the extended mind are not question-begging given their predication on a liberal form of functionalism. Adjudicating between the extended mind and its critics, therefore, seems to require adjudicating between liberal and chauvinistic forms of functionalism. But this is a dispute that has been ongoing almost since functionalism's inception. In the absence of any satisfactory resolution of this dispute, the clear danger for the extended mind is one of stalemate. (pp.6-7)

Perhaps this deadlock can be broken – and for reasons that we have met already. Recall from chapter 3 that Rupert's example of the critical divergence between cases of inner memory and cases of alleged extended memory turns on the failure of certain extended systems to exhibit the generation effect in the right sort of way. However, as we saw then, if all else is equal, neither commonsense nor scientific psychology will refuse to count, as feats of memory, the achievements of a non-extended subject who fails to exhibit the generation-effect, but whose overall context-sensitive information storage and retrieval capacity remains otherwise intact. This gives us good reason to think that the difference between exhibiting or failing to exhibit the generation effect in the right sort of way doesn't mark the boundary between having a memory and not having one, which further suggests that there must be an explanatorily useful, generic account of memory that is broad enough to cover generation-effect and non-generation-effect cases. So although Rupert may be right that for two creatures to realize the cognitive trait of *exhibiting the generation effect in memory*, they will need to share a fine-grained inner profile which resists any extended realization, that fact, if it is one, poses no real threat to ExC. Extended systems of context-sensitive information storage and retrieval that fail to exhibit the generation effect might still count as memory, and thus as cognitive.

Let's now not only generalize this result (in the way that Rupert does for his negative outcome), but also place it in the functionalist context introduced by Rowlands' recent analysis. The more general message is that exhibiting or failing to exhibit *fine-grained functional traits* (like the generation effect) doesn't mark the boundary between being a cognizer and not being one. Rather, the level of functional grain that matters for the presence or absence of cognition must be set high enough so that, other things being equal, a system that exhibits some fine-grained functional trait and one that doesn't both count as cognitive. (For additional considerations which point in the same direction, see Sprevak (manuscript, especially p.11). More from Sprevak in a moment.) In the end, then, it looks as if the Rowlands deadlock may be broken, on the grounds that the fine-grained, chauvinistic form of functionalism assumed by Rupert should be rejected in favour of a higher-level, liberal grain of functional analysis. Such a state of affairs would lend support to extended functionalism.

This looks like a tidy result. However, at the risk of undermining my own argument, I have to confess that I might have been moving just a tad too quickly. For the truth is that my own treatment of extended functionalism suggests that, *in some instances* (although not in the case of the generation-effect), *fine-grained functional differences might determine the cognitive-noncognitive*

boundary. Recall that the distinctive tenet of Clark-style microfunctionalism is that the functional roles that fix the domain of the cognitive will often be specified in terms of relations that do not respect the boundaries of linguistic or conceptual thought. It is natural to hear this as an appeal to fine-grained functional roles. For example, it is at least arguable that any architecture deserving of the title 'cognitive' will need to display capacities such as flexible generalization and the graceful degradation of performance in the face of restricted damage or noisy/inaccurate input information. Such flexibility and robustness (more on which in chapter 10) are plausibly at work in the entire suite of cognitive activities, from online perceptually guided action to offline reflection and reason. Indeed, one major impetus to the rebirth of connectionism in the 1980s was that while such capacities are often missing from, or difficult to achieve in, classical systems, connectionist networks seem to exhibit them as 'natural' by-products of their basic organization. Moreover, it is plausible that the fact that connectionist networks possess these cognition-fixing properties is explained by their microfunctionalist mode of organization. Thus, as Smolensky (1988) famously stressed, a certain kind of sensitivity to subtle shades of context-dependent meaning falls out of the common connectionist commitment to the fine-grained, bottom-up determination of conceptual-level content by way of representational schemes in which even semantic microfeatures may be realized by shifting patterns of network activation (for discussion, see Clark 1989, pp.118-121). This suggests that the realization of certain fine-grained functional roles might be crucial to cognition. So *if* it were the case that such roles could *only* be implemented internally, then we would have a case against extended functionalism.

The good news for the ExC theorist is that proving the antecedent of the critical final inference looks to be a big ask. As mentioned earlier, the microfunctionalist commitment to multiple realizability means that there is every reason to think that at least some microfunctions will be apt for realization in extended substrates. Thus imagine that Otto's notebook (see chapter 2) is not a repository of static text, but a mobile computing device armed with connectionist software capable of the sort of flexible generalization and graceful degradation characteristic of such systems. And let's assume, for the sake or argument, that the computing device (just like the notebook, according to Clark and Chalmers' original example) contributes to Otto's behaviour-generation processes in such a way that we are happy to include it as part of Otto's cognitive systems. In this case, the microfunctions that underlie the key properties of flexible generalization and graceful degradation are at least partly realized beyond the skin. What this indicates is that, in the end, the question of the grain at which

functional analysis should be performed is pretty much orthogonal to the issue of cognitive extension. In other words, the situation is not that for ExC to be true, *all* cognitive traits would need to be specified at a high level of grain, meaning that the ExC theorist assumes a liberal form of functionalism, while for ExC to be false, *all* cognitive traits would need to be specified at a fine level of grain, meaning that the opponent of ExC assumes a chauvinistic form of functionalism. Indeed, it is entirely possible that *some* of the functional roles that will be identified by an AI-centred cognitive science as determinative of cognition will be fixed at a fine level of grain. Some of these will allow for extended realizations, some may not.

5.7 …are a Million to One They Said

Our third argument that targets extended functionalism is due to Mark Sprevak (manuscript). While it shares certain features with Rowlands' argument, it demands attention in its own right. It turns on an independently plausible principle that Sprevak calls *the Martian intuition*.

> The Martian intuition is that it is possible for a creature with mental states to exist even if such a creature has a different physical and biological makeup from ourselves. An intelligent organism might have green slime instead of neurons, and it might have different kinds of connections in its "nervous" system. The Martian intuition applies to fine-grained psychology as well as physiology: there is no reason why a Martian should have exactly the same fine-grained psychology as ours. A Martian's pain response may not decay in exactly the same way as ours; its learning profiles and reaction times may not exactly match ours; the typical causes and effects of its mental states may not be exactly the same as ours; even the large-scale functional relationships between the Martian's cognitive systems (e.g. between its memory and perception) may not exactly match ours. (Sprevak, manuscript, pp.5-6)

As indicated by our previous discussion of the place of functionalism in the history of philosophy of mind, one of the key properties of that thesis (as standardly conceived) is that it gives us the conceptual resources to save the Martian intuition. However, Sprevak argues that it can achieve this only if the level of functional grain is set at a sufficiently coarse level. If the level of functional grain is set too finely, Martians whose pain responses decayed differently to ours or whose learning profiles and reaction times did not exactly

match ours would be illegitimately excluded from being cognizers, and the Martian intuition would be violated. So how does the Martian intuition bear on the case for cognitive extension? Sprevak's claim (ibid. p.8) is that "if the grain parameter is set at least coarse enough to allow for intelligent Martians, then it also allows many cases of extended cognition". Why think this? As Sprevak explains (partially echoing an argument from Clark forthcoming b – see chapter 3 above), if we take some putative case of extended cognition, such as the Otto-notebook system, we can always imagine a functionally equivalent system (e.g. Clark's Martian bit-map memory system that we have met previously) that is located entirely inside the head of a Martian. On the strength of the Martian intuition, we would count that Martian-internal system as cognitive, so when, as functionalists, we fix the level of grain for our analysis, it must be set coarsely enough to generate that result. But if it is that coarse, then the (by hypothesis) functionally identical extended system too will count as cognitive. Or at least it will do so, if we accept the parity principle. For of course it would be inner chauvinism to exclude the extended system simply because it involves external factors, when in all other relevant respects it is equivalent.

It is at this point that the trouble for extended functionalism starts. For Sprevak argues that once the level of functional grain is set coarsely enough to save the Martian intuition, what is entailed is a radical form of ExC that is wildly over-permissive. To adopt the terminology that I have used previously, the resulting account will be excessively liberal, in that it will welcome in to the domain of the cognitive certain unwanted interlopers. For example, Sprevak argues that, according to this form of ExC, if I have a desktop computer which contains a program for calculating the dates of the Mayan calendar 5,000 years into the future, then, even if I never run this program, I possess an extended cognitive process that is capable of calculating the dates of the Mayan calendar. Why? Because one could imagine a Martian with an *internal* process that is capable of calculating the dates of the Mayan calendar *using the same algorithm as my desktop computer*. Even if the Martian never has cause to use this process, nevertheless it seems right to say that it is part of that creature's cognitive architecture. Now we simply apply the parity principle: there is a functional equality between the dispositional contribution of the Martian's inner process to the Martian's behavioural repertoire and the dispositional contribution of the external desktop process to my behavioural repertoire. Since the Martian's inner process counts as cognitive, equal treatment demands that the same status be granted to the process in my desktop computer. And intuitively that seems wrong. Surely the desktop process is a potential aid to cognition, but is not itself part of my cognitive architecture.

This is bad news for extended functionalism, since if Sprevak is right, functionalism entails a wildly over-permissive form of ExC that looks to be false. But it is also bad news for functionalism as a theory of mind, since if functionalism entails a false theory, then functionalism too is false. Of course, the critical argument could be blocked if we gave up on the Martian intuition, since then, to return to Sprevak's Mayan Calendar example, the Martian inner process wouldn't count as cognitive. But that is ruled out because the Martian intuition is independently plausible. Alternatively, the critical argument could be blocked if we gave up on the parity principle, since then we could count the Martian inner process as cognitive, while denying that status to the desktop process. But that is ruled out because the parity principle is one of the keystones of the case for ExC (Sprevak manuscript, p.16). So it seems that Sprevak has created a serious dilemma for the extended functionalist who favours a parity-driven case for ExC.

Or has he? Let's look again at the structure of Sprevak's argument. The conceptual backdrop against which it operates involves three factors: a functionalist understanding of ExC, the independent plausibility of the Martian intuition, and the centrality of the parity principle to the positive case for ExC. The path to the apparently troublesome dilemma then has four steps. At step 1 Sprevak describes a distributed, environment-involving system that intuitively looks to be a wildly unlikely case of extended cognition, so unlikely in fact that any theory according to which the system as a whole counted as cognitive would, by virtue of that fact, look to be false. At step 2 he imagines a functionally identical system located entirely inside the head of a Martian, and concludes, on the grounds of a functionalism committed to the Martian intuition, that we would grant that system cognitive status and thus that the level of functional grain should be set coarsely enough to deliver that result. At step 3 he argues, on the strength of the parity principle, that the distributed system described at step 1 must also count as wholly cognitive. At step 4 he draws the anti-ExC and anti-functionalist conclusions. It's compelling stuff. So what has gone wrong?

It seems that step 2 of Sprevak's argument depends on a form of the Martian intuition that is *significantly more radical* than the one he explicitly formulates as part of his conceptual backdrop. And whereas the latter intuition does indeed command considerable plausibility, the former doesn't. To explain: What Sprevak does at step 2 is take what he assumes to be the noncognitive, externally located elements in a distributed process, place them inside the head of a Martian, and conclude that they now deserve to be rewarded with cognitive status. But where is the justification for suddenly counting these elements as

themselves cognitive? Apart from their spatial location, nothing about them has changed from when they were judged to be noncognitive. The only new factor is their recently acquired in-the-head-ness. So it certainly looks as if an external element that we took to be noncognitive has since become cognitive, *purely in virtue of being moved inside the head*. Now, the core of the Martian intuition, as explicitly formulated by Sprevak, is that "it is possible for a creature with mental states to exist even if such a creature has a different physical and biological makeup from ourselves". But it certainly doesn't follow from this highly plausible principle that any state or process that happens to be found inside the head of an intelligent Martian must, simply because of its in-the-head-ness, count as a cognitive state or process. The latter claim, which is what Sprevak seems to need for his anti-ExC argument, would constitute a significantly more radical form of the Martian intuition. Moreover, it is one that clashes unhelpfully with the parity principle that Sprevak assumes at step 3 of his argument. Indeed, it is a corollary of the parity principle that the smuggled-in, more radical form of the Martian intuition cannot be right. After all, the parity principle implies that an in-the-head element that we take to be cognitive doesn't become noncognitive purely in virtue of being moved outside the skin. And the direction of travel here is irrelevant. The more general slogan is *equal treatment regardless of location*. Thus the parity principle also implies that an external element that we take to be noncognitive doesn't become cognitive *purely in virtue of being shifted inside the head*.

What this suggests is that the extended functionalist can avoid Sprevak's dilemma by refusing to endorse the more radical form of the Martian intuition. This is something that the fan of the parity-driven case for cognitive extension ought to do anyway, given that the parity principle is inconsistent with that version of the intuition. The orthodox version, the one explicitly stated by Sprevak, remains in force, of course. But that is consistent with the claim that the class of Martian in-the-head elements (indeed, the class of in-the-head elements in general) may contain some noncognitive members, and so does not entail that in-the-head elements whose contribution to intelligent behaviour is functionally identical to that of certain noncognitive external elements attain cognitive status purely in virtue of becoming intra-cranial. It is also fully compatible with the parity principle. The path to Sprevak's dilemma is thus blocked, at step 2. The missing piece of the puzzle here is our old friend, a locationally independent account of the cognitive that fixes the benchmark for parity. Once such an account is part of our conceptual picture, there is no reason to think that any old process will count as cognitive, just because it has been rammed inside the head of a Martian. The benchmark does sterling theoretical work in weeding out

unwanted interlopers into the domain of the cognitive, wherever they happen to be located.[4]

It might seem that my response to Sprevak's argument commits the extended functionalist to saying something uncomfortable about the Mayan calendar case. To see why, let's say we begin our consideration of the issues simply by imagining a Martian who has an inner program capable of calculating the dates of the Mayan calendar 5,000 years into the future. Even though, by hypothesis, this piece of inner machinery is never actually used, it might seem that we should have no misgivings about awarding it cognitive status. On the other hand, my treatment of the full Sprevak argument suggests that this is a mistake. The point, then, is that if we begin by focussing on the inner Martian program, rather than the one stored on my desktop computer, the refusal to count the first instance as a cognitive mechanism looks to be much less well motivated. What I think has happened here is that we have been seduced by some subtle and unnoticed changes to the details of the hypothetical scenario, changes that have been surreptitiously introduced by the variation in set-up. When we begin our consideration of the issues by imagining the Martian inner program, our natural tendency is to think of that mechanism as being already functionally integrated into (although not yet activated within) an organized economy of states and processes. Those states and processes are intimately embedded in subtle and complex perceptual, memory and reasoning systems that have been evolved or developed in relation to each other, and that already meet whatever the criteria are for cognitive status. If the desktop program for calculating the Mayan calendar were an element in this kind of functional economy, then it may seem far less crazy to conclude that it may be a cognitive mechanism, or at least part of one, even though it is spatially located outside the head. Various factors might pump our intuitions in this direction. Perhaps the program is configured to reflect a particular individual's favoured kind of interface, and has been made remotely accessible through real-time mobile computing technology or will, in the future, be made available at the firing of a neuron through a brain implant that connects the mechanism to a wireless network. Never mind the cyborg imagery. However one develops the basic idea, the resulting image is a long way from the one suggested by Sprevak's decoupled desktop. For when we begin our consideration of the issues by focussing on the desktop program, our natural tendency is to think of a stand-alone and removable software application, sitting on a machine that sometimes achieves some fancy feats of text-editing, graphics, and information retrieval, but which, in the end, is no more than a sophisticated tool for work or play. This encourages us to find it wildly unlikely that the program in question could ever count as a cognitive

mechanism, even if it were to be transported inside a Martian head. In the end, then, the presence of certain subtle differences between how things are being set-up in our two cases means that Sprevak's dilemma can still be held at bay.

What has been argued in this chapter so far is that one immensely profitable way to unpack ExC is as a generically functionalist thesis. What has not been argued is that ExC is *necessarily* functionalist in form. One could begin to make conceptual room for a functionalism-free ExC, *if* one could make conceptual room for functionalism-free multiple realizability. Churchland (2005) has claimed that the latter is possible. Adopting a nonequilibrium thermodynamic framework for understanding psychological phenomena, Churchland suggests that cognitive creatures – by which he means creatures that learn about the world – should be conceptualized as 'extra-somatic information multipliers'. By exploiting information which they already possess, such creatures come to embody additional information about their environments, usually through progressive neural rewiring. Churchland then proceeds to unpack this broad vision of cognition in recognizably connectionist terms, according to a theoretical framework in which activation vectors and transformations between those vectors are held to be regulated by adjustable weighted connections tuned by learning. For our purposes, the key move in Churchland's analysis comes next, with the claim that although the vector processing dynamics just highlighted may be realized in a range of material substrates (mammalian brains, octopus brains, electronic chips, and so on), they are not functionalist in form. Thus we arrive at the conclusion that functionalism is not necessary for multiple realizability. But now what supports the final anti-functionalist move? The answer is that, in Churchland's thinking, functionalism is fettered to the conceptual-level categories of folk psychology. Thus he identifies a central claim of classical functionalism as being that "Folk Psychology – our common-sense conception of the causal structure of cognitive activity – already embodies a crude and partial representation of the function we are all (more or less) computing" (Churchland 2005, p.3 manuscript). For Churchland, then, the fate of functionalism is determined by the fate of folk psychology. And this is bad news for functionalism, because the vector processing dynamics described above are unlikely to respect the boundaries of our higher-level folk-psychological categories, meaning that folk psychology is under threat of elimination. But if folk psychology goes, so does functionalism. Thus we dispose of functionalism while leaving multiple realizability intact.

How might we resist Churchland's argument? As it happens, one means to do so is already within our grasp. For although it may well be true of classical

functionalism that its categories are intertwined with those of folk psychology, the same surely isn't true of the Clark-style microfunctionalism that we encountered in the last section. After all, the defining characteristic of microfunctionalism is its claim that the functional roles that matter for cognition will often be specified in terms of fine-grained relations that do *not* respect the boundaries of linguistic or conceptual thought. So while Churchland's connectionist-friendly vision of cognition as vector processing is plausibly at odds with classical functionalism, it's hard to believe that it isn't entirely compatible with Clark's connectionist-friendly microfunctionalism. But if this is right, then even if folk-psychology and classical functionalism are ripe for elimination, microfunctionalism is still a going concern. And that form of functionalism may yet provide the conceptual resources by which the multiple-realizability of cognitive phenomena is to be secured. In the end, then, Churchland's argument fails to establish that one can have multiple realizability without functionalism.

Still, even though Churchland's specific argument is flawed, it really does seem that it ought to be possible to gain conceptual access to the ExC-crucial property of multiple realizability without endorsing functionalism. Indeed, although we ultimately ruled out the raw information processing account of cognition, on grounds of excessive liberality, one of the initial attractions of that approach was that it was locationally uncommitted, a phenomenon which, as we know, is intimately tied up with multiple realizability. What this suggests is that non-functionalist versions of ExC are potentially available. Of course, if I am right, the functionalist perspective ought to be attractive to the ExC theorist because, understood a certain way, that perspective plays an important part in supplying a principled answer to the pivotal question, for ExC, of what fixes the benchmark for parity. Moreover, I have argued that it is possible to resist certain arguments that would, if successful, have undermined the extended functionalist point of view. On the other hand, it is common philosophical knowledge that functionalism *as a general theory of mind* faces some serious difficulties, especially, perhaps, in the area of phenomenal consciousness – the what it's-like-ness of experience. Who can forget philosophical evergreens such as the single system comprising the entire Chinese nation, organized so as to satisfy the functional definition of a mind (Block 1980), or the functionally-identical-to-one-of-us zombie (Chalmers 1996). In both cases the message is supposed to be that since we enjoy phenomenal consciousness, yet certain systems functionally identical to us plausibly don't, no purely functional characterization can explain phenomenal consciousness. This result, if correct, would leave functionalism sorely lacking as a theory of mind. (I shall come back

to some of these general anti-functionalist worries in chapter 8.) Anyone who is moved by such cases to harbour doubts about functionalism in general, but who is attracted by ExC, might well be moved to seek out any viable non-functionalist options. This suggests that we have a final job to do before we finally declare parity-based extended functionalism to be the best available development of ExC. We need to see what the literature has to offer in the way of further alternatives. This task will be postponed until chapter 8. Ahead of that it is finally time to broaden the scope of our discussion, and to see how the joints of biological nature are being recarved in other ways.

Notes

1. The claim that ExC is in some way a form of, dependent on, or at least most naturally, commonly or profitably played out in terms of, functionalism is now pretty much part of the received view of things; see e.g. Adams and Aizawa (2008), Clark and Chalmers (1998), Clark (2005, 2008a, 2008b, forthcoming a, b), Menary (2007), Rupert (2004), Sprevak (manuscript), Wheeler (2008).

2. In chapters 3 and 4, we encountered an argument against ExC that depended on the claim that because some psychological property attracts the attention of cognitive psychologists, that property must in some way play a role in marking off the cognitive from the non-cognitive. I argued against that thought. At the very least it seems clear that from the fact that some specified phenomenon happens to be of interest to cognitive psychologists, one cannot infer that failing to exhibit that phenomenon is an indication of non-cognitive status. However, notice that once our benchmarking machinery is sourced not from an orthodox science of human cognition, but from a science of cognition *in general*, we *do* have reason to think that phenomena that fall outside its domain of interest also fall outside the domain of the cognitive. After all, as we have seen, the explanatory scope of such a science extends to all possible minds.

3. As I have discussed at length previously (e.g. Wheeler 1994; 1998; 2005a; 2005b), most work in connectionist cognitive science has tended to concentrate on network architectures that, in effect, limit the range and complexity of the dynamics available to such a system, when compared with their biological cousins. Such 'abiological' restricting features include: neat symmetrical connectivity; noise-free processing; update properties which are based either on a global, digital pseudo-clock or on methods of stochastic change; units which are uniform in structure and function; activation passes that proceed in an

orderly feed-forward fashion; and a model of neurotransmission in which the effect of one neuron's activity on that of a connected neuron will simply be either excitatory or inhibitory, and will be mediated by a simple point-to-point signalling process. We shall return to this issue in later chapters.

4. For those with long-ish memories, this might seem reminiscent of Adams and Aizawa's refusal to grant the Martian bit-map system cognitive status, on the grounds that it failed to meet their favoured mark of the cognitive (see chapter 4). What my response to Sprevak's argument indicates is that, in my view, there was nothing fundamentally wrong with Adams and Aizawa's basic thought that some in-the-head elements may be vehicles for exclusively noncognitive processes. The difficulty for Adams and Aizawa, as I explained at the time, is with how to secure that move in the case of the Martian bit-map system, given what they take the mark of the cognitive to be (a combination of involving non-derived representations and of being individuated by certain information processing mechanisms identified by human cognitive psychology).