In R. Menary (ed.), *The Extended Mind*, MIT Press, Cambridge, Mass.. 2010, pp.245-270.

# In Defense of Extended Functionalism

**Michael Wheeler**

## 1. The Dynamic Duo

According to the extended cognition hypothesis (henceforth ExC), there are conditions under which thinking and thoughts (or more precisely, the material vehicles that realize thinking and thoughts) are spatially distributed over brain, body and world, in such a way that the external (beyond-the-skin) factors concerned are rightly accorded fully-paid-up cognitive status.[1] According to functionalism in the philosophy of mind, "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part" (Levin 2008). The respective fates of these two positions may not be independent of each other. The claim that ExC is in some way a form of, dependent on, entailed by, or at least commonly 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, 2008, this volume a, b, forthcoming; Menary 2007; Rupert 2004; Sprevak manuscript; Wheeler forthcoming). Thus ExC might be mandated by the existence of functionally specified cognitive systems whose boundaries are located partly outside the skin. This is the position that Andy Clark has recently dubbed *extended functionalism* (Clark 2008, forthcoming; see also Wheeler forthcoming).

Against this background, the present paper has two main goals. The first (sections 2 and 3) is to clarify and amplify the relationship between ExC and functionalism, and thereby to plot the path to extended functionalism. The second (sections 4, 5 and 7) is to defend extended functionalism against three potentially damaging critical assaults. Section 6 is an interlude that highlights a key aspect of the extended functionalist picture. The paper ends (section 8) with a brief (and I mean brief) remark on extended functionalism and phenomenal consciousness.

## 2. The Extended Cognition Hypothesis

ExC is a view about the whereabouts of thinking and thoughts that is distinct not only from the position adopted by orthodox (classical or connectionist) cognitive science, but also from the position adopted by any merely embodied-embedded account of mind. That is why my opening characterization of ExC included the qualification that the target phenomena must be distributed over brain, body and world, *in such a way* that the external (beyond-the-skin) factors concerned are themselves rightly accorded fully-paid-up cognitive status. In other words, as Adams and Aizawa (e.g. 2008, this volume) have repeatedly emphasized, it is not sufficient for genuine cognitive extension that thinking be spatially distributed over brain, body and world solely in the weak sense that applies when some instance of intelligent behaviour is discovered to be causally dependent, perhaps in previously unexpected ways, on the bodily exploitation of certain external props or scaffolds. We may even introduce the additional feature that the cognitive task in question could not have been achieved by brains like ours without the causal contribution of the external elements in question. Still the shortfall remains. Bare causal dependence of mentality on external factors – even when that causal dependence is of the 'necessary' kind just highlighted – is simply not enough for genuine cognitive extension. What is needed is the *constitutive* dependence of mentality on external factors, the sort of dependence indicated by talk of the beyond-the-skin factors themselves rightly being accorded fully-paid-up cognitive status. Only this latter kind of distribution – we might call it *ontological distribution* – will do.

In order to illustrate this crucial point, we can adapt an analysis due originally to Rumelhart et al. (1986) that has since become something of a stock example in the embodied-embedded-extended mind literature. Most of us solve difficult multiplication problems using pen and paper.[2] The pen and paper resource is a beyond-the-skin factor that helps to transform a difficult cognitive problem into a set of simpler ones and acts as a temporary store for the results of intermediate calculations. For orthodox cognitive scientists *and for supporters of the merely embodied-embedded view of mind*, the pen and paper system is to be conceived as a noncognitive environmental prop. It is an external tool that aids certain cognitive processes via embodied interaction, but is not itself a proper part of those processes. Of course, orthodox cognitive scientists and embodied-embedded theorists differ on how best to characterize the interactive arrangement of skin-side cognitive processes and external prop. In particular, the embodied-embedded theorist is likely to count the bodily activity involved as itself a cognitive process, as opposed to a mere output of neurally located

cognition, and to trace rather less of the source of the manifest complexity of the observed behaviour to the brain, and rather more to the structured embodied interactions with the external pen and paper system. For all that, however, both of these camps ultimately think of cognition as a resolutely skin-side phenomenon. By contrast, the ExC theorist considers the coupled combination of pen-and-paper resource, appropriate bodily manipulations, and in-the-head processing to be a cognitive system in its own right, a system in which although the differently located elements make different causal contributions to the production of the observed intelligent activity, nevertheless each of those contributions enjoys a *fully cognitive* status. In my view, the supporting case for the hypothesis of embodied-embedded cognition has been successfully made over and over again.[3] If that's right, then the key issue facing ExC theorists right now is not how to argue against the received (if that's what it still is) orthodox view in cognitive science, but rather how to justify the transition from a 'merely' embodied-embedded mind to an extended one.[4]

**3. From Functionalism to Extended Functionalism**

Some of the conceptual machinery required to effect the transition just identified plausibly comes in the form of a familiar philosophical theory of mind, namely *functionalism*. According to the traditional formulation of this view, the canonical statement of which is arguably due to Putnam (1967), a mental state counts as the mental state it does because of the causal relations it bears to sensory inputs, behavioural outputs, and other mental states. Who gets to decide what the psychologically relevant causal relations are (e.g. philosophers performing conceptual analyses of folk-psychological terms, psychologists performing scientific experiments) is a matter of intellectual debate. For the present the key point is this. As every undergraduate who has ever taken a class in philosophy of mind knows, traditional functionalism triumphantly frees us from a kind of neural or carbon chauvinism about the mind. In so doing it bolsters the intellectual credentials of Doctor Who, Star Wars, Ben 10, and every other science fiction adventure predicated on encounters with alien intelligence. It also keeps the good people of SETI in their jobs. In other words, traditional functionalism provides a principled basis for concluding that creatures whose brains happen to be built out of physical stuff different from our own may still be cognizers. It achieves this heady feat because it bequeaths to the mind the chauvinism-busting property of *multiple realizability*. To explain: if psychological phenomena are constituted by their causal-functional roles, then our terms for mental states, mental processes, and so on pick out equivalence classes of

different material substrates, any one of which might in principle realize the type-identified state or process in question. But of course that means that robots, Martians and the Ood and may all join us in having mental states, just so long as the physical stuff out of which they are made is capable of being organized so as to implement the right functional profiles.

What has this brief excursion into the history of philosophy got to do with ExC? The answer, I suggest, is that one of the standard considerations used in pro-ExC arguments, namely *the parity principle*, forges a strong connection between functionalism and ExC. To remind us of the parity principle, here is a much-quoted passage from Clark and Chalmers (1998, p.8). "If, as we confront some task, a part of the world functions as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head." In broad terms, then, the parity principle states that if there is functional equality with respect to governing behaviour, between the causal contribution of certain internal elements and the causal contribution of certain external elements, and if the internal elements concerned qualify as the proper parts of a cognitive trait, then there is no good reason to deny equivalent status – that is, cognitive status – to the relevant external elements. Parity of causal contribution mandates parity of status with respect to inclusion in the domain of the cognitive.[5]

So what? The parity principle is based on the thought that it is possible for the very same type-identified cognitive state or process to be available in two different generic formats – one non-extended and one extended. Thus, in principle at least, that state or process must be realizable in either 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. So, if we are to argue for cognitive extension *by way of parity considerations*, the idea that cognitive states and processes are multiply realizable must make sense. Now, as we have seen, functionalism provides one well-established platform for securing multiple realizability. That said, we don't quite have a case of plug and play philosophy here. Functionalism – or rather how we formulate it – needs to be tweaked a little before current needs are met. To see why recall that, according to the traditional formulation of the position as given earlier, a mental state is constituted by the causal relations that it bears to sensory inputs, behavioural outputs, and other mental states. But depending on how one hears terms like 'sensory inputs' and 'behavioural outputs', this statement of the view may harbour a bias towards the inner that isn't, at root, a feature of its 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 (cf. the formulation given by Levin 2008, as quoted near the beginning of this paper). Once we give this more general characterization 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 partly outside the skin. It is in this way that we arrive straightforwardly at the position that, following Clark, I shall call *extended functionalism* (Clark 2008, forthcoming; see also Wheeler forthcoming). I think that extended functionalism is an attractive position with good philosophical and cognitive-scientific credentials. Not everyone agrees.

## 4. Troubles for Extended Functionalism Part I: the Adams-Aizawa Distinctiveness Principle

As part of their sustained critical treatment of ExC, Fred Adams and Ken Aizawa (2008) 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 allegedly 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). Call this the *Adams-Aizawa Distinctiveness Principle*. As evidence for the way in which this principle plausibly identifies neuronal states and processes as the only vehicles of cognition, Adams and Aizawa point to the 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.

It is at this point 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). As far as I can tell, the specific language of 'processes' is not essential to Adams and Aizawa's point, which ultimately concerns the distinctiveness of a range of relevant phenomena (including for example states and mechanisms, as well as processes) at the different levels. With that clarification in place, we can see that Adams and Aizawa's argument implies a rejection of the general claim that human cognitive traits are sometimes multiply realized. To be clear: Adams and Aizawa do not reject the *in-principle* possibility of cognition-realizing substrates that involve (wholly or partly) non-neuronal elements. What they reject is the idea that minds like ours are *in fact* ever realized by such substrates. Understood as part of their general critique of cognitive extension, now interpreted in terms of extended functionalism, their argument is thus levelled not against the in-principle possibility of cognitive extension, but against the idea that minds like ours are in fact ever extended. In view of all this, one defensive strategy open to the ExC theorist would be to find examples of scientifically well-established cases which show that the Adams-Aizawa distinctiveness principle is false. If there are extant distinctive higher-level phenomena, such that each of those phenomena is, *in fact*, multiply realized by more than one kind of lower-level phenomenon, then we would have no *general* reason to expect each distinctive higher-level phenomenon to be realized exclusively in a single material substrate, and thus no *general* reason to expect cognition in particular to be realized exclusively in a neuronal substrate.

As it happens, it seems that the evidence needed by the ExC theorist is plentiful, in examples of what is known in biology 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 a distinctive higher-level phenomenon (relatively speaking) that is multiply realized. The Adams-Aizawa distinctiveness principle is false. Extended functionalist minds may yet be actual.

**5. Troubles for Extended Functionalism Part II: the Rowlands Deadlock**

A second analysis that, in a different way, questions the ability of extended functionalism to deliver cognitive extension hails from Mark Rowlands (manuscript). According to Rowlands, if one reflects on the interplay between (i) an argument against parity-driven ExC developed by Rob Rupert (2004; for related considerations see Adams and Aizawa 2008) and (ii) a way of responding to Rupert's argument that I have been known to pursue (Wheeler forthcoming), what emerges is a deadlock between the two sides, the paralyzing character of which may be traced to the functionalist terms of the debate. In what follows I shall lay down a path that leads to this stalemate, a path that adds detail to Rowlands' own analysis, but which ends up at the same unfortunate (for ExC) point. Let's begin, then, by revisiting Rupert's argument against ExC and what might be wrong with it.

Rupert calls on empirical psychological data which, he argues, may be used to indicate significant differences between the profile of internal memory and the profile of certain external resources, as such external resources might plausibly figure in the process of remembering. According to Rupert, such differences tell against any attempt to see the latter phenomena as being of the same explanatory kind as the former. For example, there are psychological experiments which show that internal memory is sensitive to what is called the generation effect. Where this effect is in evidence, subjects gain a mnemonic advantage by generating their own meaningful connections between paired associate items be learned. Rupert argues that the generation effect will simply not occur in some extended 'memory' systems (e.g., in a system according to which, during recall, the subject refers to a notebook in which the paired associates are accompanied by connection sentences produced by those subjects during learning, but which were entered into the notebook by the experimenter). He concedes that it might occur in others (e.g., in a system according to which, during recall, the subject refers to a notebook in which the paired associates to be learned are accompanied by connection sentences produced and entered by the subjects during learning). In the latter case, however, he suggests that the effect is an accidental feature, rather than an essential or definitional dimension, of the memory system. Rupert concludes

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. The final step is to generalize from this conclusion about memory to a conclusion about all cognitive traits. As Rupert points out, this step is plausibly justified by the fact that memory is a core cognitive trait, suggesting that what goes for memory goes for cognition in general.[6]

Rupert's argument has the following form: first we identify certain features of some core cognitive trait as standardly (internally) conceived that are not shared (or not shared in the 'right' way) by any extended arrangement that might be thought to perform the same cognitive task; then we conclude that since the parity principle is not satisfied, ExC is false. But once this two-part structure is exposed, the parity-driven ExC theorist will want to lodge a complaint (Wheeler forthcoming). For although in general that theorist must concede the existence of the kinds of functional differences identified by Rupert, she will want to object to the further claim that such differences result in a breakdown of parity. What allows the ExC theorist to block this further claim is the fact that it depends on a seemingly contestable assumption that the benchmark for parity (in effect, what counts as cognitive) should be set by the extant fine-grained details of the human inner. It is only because these details are being allowed to call the cognitive shots that the divergent functional profiles exhibited by the extended systems in question mandate the judgment that those systems should be denied cognitive status. However, when properly understood, the parity principle does not privilege the organization and processing of the actual human inner in the way that Rupert's argument suggests. Full discussion of this issue would take us too far afield (for an extended treatment, see Wheeler manuscript). But, in somewhat sketchy and general terms, here is a way of unpacking the appeal to parity so that ExC is insulated against Rupert's concerns. First we give an account of what it is to be a proper part of a cognitive system that is fundamentally independent of where any candidate element happens to be spatially located. Then we look to see where cognition falls – in the brain, in the non-neural body, in the environment, or, as the ExC theorist predicts may sometimes be the case, in a system that extends across all of these aspects of the world. On this model, parity is conceived not as parity with the inner simpliciter, but rather as parity with the inner *with respect to a locationally uncommitted account of the cognitive*. Although I am no legal philosopher, it seems to me that this way of understanding the notion of parity in cognitive theory has a recognizable and illuminating (although arguably slightly strained) analogue in the way that two citizens of a democratic state may be understood as having

the right to equality of treatment under the law. Ignoring cases of precedence, what counts as the correct treatment under the law is presumably not fixed by the case of one of the parity-enjoying citizens. Rather, each of the two citizens enjoys parity with the other with respect to an independently fixed standard of correct legal treatment.[7]

At this point one might wonder what remains of Clark and Chalmers' original idea that, in applying the parity principle, we should ask of some external process that plays a part in governing behaviour, '*Were this process done in the head*, would we have any hesitation in recognizing it as part of a cognitive process?'. The first thing to note here is that the appeal to the inner contained in this method for reaching a judgment regarding parity is not an appeal to the fine-grained profile of the extant human inner. All that happens in the thought experiment is this: certain external processes get shifted spatially, across the boundary of the skin, in an inwardly moving direction. Of course, we are not supposed to imagine that the relevant externally located physical elements themselves are grafted onto the brain. Rather, we imagine that exactly the same functional states and processes that are realized in the actual world by those externally located physical elements are now realized by certain internally located physical elements. Having done this, if we then judge that the now-internal but previously external processes count as part of a genuinely cognitive system, we are driven to conclude that they did so in the extended case too. After all, by hypothesis, nothing about the functional contribution of those processes to intelligent behaviour has changed. All that has been varied is their spatial location. And if one were to claim that that spatial shift alone is sufficient to result in a transition in the status of the external elements in question, from noncognitive to cognitive, one would, it seems, be guilty of begging the question against the ExC theorist. Now notice that at no point in this explanation of how the appeal to the inner contained in the parity principle works have we been forced to use the fine-grained profile of the extant human inner in order to determine what counts as cognitive. In other words, the application of the parity principle does not itself set the benchmark for parity (fix what counts as cognitive). Instead it acts as a heuristic device designed to free us from what Clark (2007, p.167) has called "the pervasive distractions of skin and skull".

Of course, given the stress that the foregoing analysis places on functional role in judgments of cognitive status, one thing that this initial response to Rupert does is re-emphasize the connection between functionalism and ExC, at least where the latter is played out by way of parity considerations. Indeed, if the critic of ExC refused to endorse a broadly functionalist theory of mind, the

aforementioned charge of question-begging would arguably lose some of its force. Without functionalism to sustain the multiple realizability of the mental, conceptual space would remain for the claim that cognitive states and processes are somehow intrinsically related to the materiality of the target system in such a way that multiple realizability fails. Given a failure of multiple realizability, the imagined inward shift across the boundary of the skin would presumably have an impact on whether the processes in question were cognitive or noncognitive in character, even if the external factors in the extended case and the relevant inner factors in the wholly inner case enjoyed functional equivalence with respect to governing intelligent behaviour. This observation points to an under-appreciated and under-explored tension between extended functionalism and any embodied cognition view which holds that human thought and experience are tied inextricably to the details of human bodily form. Given the goals of the present analysis, however, this particular conflict will not detain us here. (For preliminary investigations of the issue, see Clark 2008, forthcoming; Wheeler forthcoming.) Our concern is with a deadlock that, as we are about to see, emerges *within* a broadly functionalist framework, *between* extended and non-extended versions of that view.

What the Rupert-style critic of ExC needs to unearth is independent support for the key assumption that the benchmark for parity should be set by the extant fine-grained details of the human inner. It might be thought that Rupert himself has the resources to marshal such support, given that his appeal to the inner is supposed to be founded not on some pro-inner prejudice or some unwarranted theoretical conservatism, but rather on a healthy and entirely defensible respect for the methods and results of contemporary cognitive science. Thus he writes: "[a]s cognitive science currently describes its explanatory kinds, they are not likely to have realizations with external components. If, for example, cognitive science is to characterize functionally the causal role of memories, this characterization must be tailored to accommodate the generation-effect, various forms of interference, the power laws of learning and forgetting and the rest" (Rupert 2004, pp.423-4; for similar reasoning, see Adams and Aizawa 2008, pp.140-1). Two aspects of this short quotation are crucial. The first is that Rupert takes current cognitive science to be a broadly *functionalist* enterprise (its job being to "characterize functionally" psychological phenomena). The second is that, by 'cognitive science', Rupert means *conventional human-oriented and inner-oriented cognitive psychology* (note the list of psychological phenomena that Rupert gives at the end of his quotation). What this tells us is that the justification for the assumption that the benchmark for parity should be set by the extant fine-grained details of the inner comes from the idea that what counts

as cognitive should be fixed by the details of the functional organization of human cognition, as identified by conventional human-oriented and inner-oriented cognitive psychology. In effect, then, Rupert is arguing for a *chauvinistic* form of functionalism that privileges the scientifically identified human-specific inner. But the extended functionalist is unlikely to be moved by this extra consideration. Why, she will ask, should we privilege conventional human-oriented and inner-oriented cognitive psychology in this way? Indeed, it seems that Rupert's more developed argument continues to beg the question against extended functionalism. For, as we have seen, extended functionalism looks to be predicated on the more liberal form of functionalism that generates a locationally uncommitted account of the cognitive.

It is at this point in the exchange of argument and counter-argument that the problem highlighted by Rowlands emerges. Here it is, in Rowlands' own words:

> This charge [that Rupert's objections are question-begging] has been leveled by Wheeler ([forthcoming]). However, this charge seems to cut both ways. 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. (Rowlands manuscript, pp.6-7)

If this problem is genuine, it spells bad news for extended functionalism. For if Rowlands is right, then to the extent that ExC is allied to functionalism, the best it can achieve against its critics is a stalemate. This is what I shall call the *Rowlands deadlock*.

Is there a way out of the impasse – one that ultimately finds in favour of ExC? Perhaps there is. Imagine we came across a human being whose purely inner memory system didn't exhibit the generation effect, but who nevertheless continued to achieve the context-sensitive selective storage and retrieval of information. I for one have no doubt at all that conventional human-oriented cognitive psychologists would find the functional difference between this generation-effect-free subject and normal human subjects extremely interesting,

and that those same psychologists would use their well-honed experimental protocols to probe and explain that difference. But I cannot conceive of any cognitive psychologist concluding that the latter subject lacks the cognitive trait of memory. So why think that exhibiting the generation effect is a defining feature of (human) memory, rather than an accidental feature? And if that's right, then what is the justification (aside from pro-inner prejudice and unwarranted conservatism) for refusing to apply the notion of memory to an extended system with a similar profile to our generation-effect-free subject? The fact that the answers to these questions are 'one shouldn't' and 'there isn't one' 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. That account will be apt to encompass, within the category of memory, extended mechanisms for context-sensitive information storage and retrieval that don't exhibit the generation effect. So although Rupert may conceivably 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.

It is clear enough that this result is not restricted to memory. Similar arguments could be developed for prediction systems that don't fall for the gambler's fallacy, inference systems that don't exhibit the patterns characteristically revealed by the Wason selection task, and so on. What our reflections suggest, then, is a general principle: just because some specified mode of functional organization happens to be of interest to cognitive psychologists, one cannot infer that the difference between exhibiting that mode of organization and not exhibiting it must in some way play a decisive role in marking off the cognitive from the noncognitive. As the case of the generation-effect-free subject indicates, such functional differences – differences that cognitive psychologists will surely want to investigate – may well be differences *within* the domain of the cognitive. The message here is not, of course, that no mode of functional organization that ever interested a cognitive psychologist could ever be relevant to the issue of how to determine membership of the cognitive. A mechanism that failed to implement the context-sensitive storage and retrieval of information simply wouldn't be memory, wherever it happened to be located. The message, rather,

is that working out whether or not a particular mode of functional organization matters to this issue will not be decided by the fact that orthodox cognitive psychologists have studied systems that exhibit it.

If we place the preceding analysis in the explicitly functionalist context that apparently generates the Rowlands deadlock, its lesson is that the difference between 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 we have ExC-independent reasons for rejecting the fine-grained, chauvinistic form of functionalism assumed by Rupert, in favour of a higher-level, liberal grain of functional analysis. Such a state of affairs paves the way for extended functionalism.

At this juncture it might seem that the Rowlands deadlock is lurking just out of sight, waiting impatiently to reappear. For although I have just offered reasons, independent of ExC, for rejecting chauvinistic functionalism in favour of liberal functionalism, so the critic of ExC might offer reasons, independent of any case against ExC, for rejecting liberal functionalism in favour of chauvinistic functionalism. For example, the critic might claim that any attempt to fix a generic functional notion of, for example, memory, one that would subsume all the relevant internal and extended systems (those that don't exhibit the generation effect, those that do, those that don't exhibit negative transfer interference effects [see note 6], those that do, and so on) would need to be so devoid of detail (in order to subsume all the different functional profiles) that it would fail to earn its explanatory keep (for this sort of argument, see, e.g., Rupert 2004). In short, the charge is that our more liberal form of functionalism is pitched at such a stratospheric level of generality that it fails to support useful psychological theorizing. And that provides a reason to favour chauvinistic functionalism. But now if there are not only appropriate and defensible reasons for adopting ExC-friendly liberal functionalism, but also equally appropriate and equally defensible reasons for adopting ExC-unfriendly chauvinistic functionalism, then the Rowlands deadlock is restored.

Once again, however, I think the stalemate can be broken. Recall yet again our hypothetical subject whose inner mechanisms of context-sensitive information storage and retrieval do not exhibit the generation effect. As we have seen, the fact that neither commonsense nor cognitive psychology baulks at the thought that this subject's feats should count as genuine cases of remembering gives us good reason to think that there must be a generic notion of what memory is that is broad enough to cover generation-effect and non-generation-effect cases. Now we can add a further observation. The fact that our subject's abilities would undoubtedly be investigated by cognitive psychologists as one possible form of the psychological phenomenon of memory surely indicates that the generic notion of memory that underwrites this way of proceeding is doing important work in organizing and shaping the project of cognitive-scientific explanation. Thus, on the strength of this example, it seems that the explanatory credentials of that generic notion of memory are in perfectly good order. And that is good news for the liberal version of functionalism that provides the theoretical backdrop against which that generic notion of memory makes sense. For it surely suggests, pace the critic of ExC, that that liberal, ExC-friendly version of functionalism is not stymied by explanatory impotence. If this is right, then the restored form of the Rowlands deadlock is ultimately unsustainable.

**6. Interlude: Extended Microfunctionalism**

So far I have been running with the thought that extended functionalism is naturally predicated on a liberal version of functionalism. Part of the supporting argument has involved the claim that the cognitive-noncognitive boundary does not coincide with the sorts of fine-grained functional differences exemplified by the difference between exhibiting or not exhibiting the generation effect. But this is not the whole story. For in spite of what I have argued so far, the fact is that the cognitive-noncognitive boundary may *sometimes* (although not in the generation-effect case) be determined by fine-grained functional differences. Here is some evidence for this conclusion. It is at least arguable that any architecture deserving of the title 'cognitive' will need to display capacities such as flexible (i.e. context-sensitive) generalization and the graceful degradation of performance in the face of restricted damage or noisy/inaccurate input information. Such capacities are plausibly at work in the entire suite of cognitive activities, from online perceptually guided action to offline reflection and reason. So how do we explain them? To reveal *part of* the answer to this question, recall that one major impetus to the rebirth of connectionist artificial intelligence (AI) in the 1980s was that while capacities such as flexible

generalization and graceful degradation are often missing from, or difficult to achieve in, classical AI systems, connectionist networks seem to exhibit them as 'natural' by-products of their basic mode of organization. So what explains this propensity? It has frequently been noted (perhaps most famously by Smolensky 1988) that the cognitively relevant functions implemented by connectionist networks will often be specified in terms of mathematical relations (between units) that do not respect the boundaries of linguistic or conceptual thought. Given the tendency (it is far from a universal commitment) of classical AI theorizing to adopt functional specifications that do respect the boundaries of linguistic or conceptual thought, one might gloss this point by saying that the salient functional roles that matter for connectionist theorizing are typically pitched a finer level of grain than those performed by classical computational systems. That's part of the reason why Clark (1989, 1999) has described connectionist theory as a kind of *microfunctionalism*. Moreover, it is highly plausible that cognitively critical properties such as flexible generalization and graceful degradation may be emergent properties of connectionist networks in part precisely because those networks are functionally organized in a fine-grained way. As Clark (1989, pp.35-6) puts it

> [Microfunctionalism] would describe at least the *internal* functional profile of the system (the internal state transitions) in terms far removed from.. contentful purposive characterizations. It would delineate formal (probably mathematical) relations between processing units in a way that when those mathematical relations obtain, the system will be capable of vast, flexible structural variability and will have the attendant emergent properties. By keeping the formal characterization… at this fine-grained level we may hope to guarantee that any instantiation of such a description provides at least potentially the right kind of substructure to support the kind of flexible, rich behavior patterns required for true understanding.

This provides evidence for the following claim: for some properties that, one might argue, would need to be displayed by any system worthy of the label 'cognitive', the fact that the system realizes a certain fine-grained functional profile may well be crucial to the possession of that property.

Of course, if it were the case that the sorts of fine-grained functional roles just highlighted could *only* be implemented internally, then this would present a serious barrier to extended functionalism. The good news for the extended

functionalist, however, is that microfunctionalism is not antithetical to the possibility of extended realizations. Significantly, as Clark (1999, p.40) notes, 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". This preservation of the functionalist commitment to multiple realizability clears the way not only to non-standard organic implementations of the microfunctions in question, as Clark's text here directly suggests, but also to extended implementations. In this context, notice that, in the longer quotation from Clark reproduced just above, he states that "*at least* the *internal* functional profile of the system would be described in microfunctionalist terms" [first emphasis mine]. In my view this way of putting the point is too conservative. There is every reason to believe that at least some microfunctions will be apt for realization in extended substrates. Thus imagine that I possess 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, just for the sake or argument, that the computing device contributes to my behaviour in such a way that, on the strength of parity-principle reasoning, we are happy to include it as part of my 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 a locationally uncommitted cognitive science as determinative of cognition will be fixed at a fine level of grain. The implication – one that enriches our vision of ExC – is that extended functionalism has a robustly microfunctionalist dimension.

## 7. Troubles for Extended Functionalism Part III: the Sprevak Dilemma

Our third argument against extended functionalism is due to Mark Sprevak (manuscript). Although this argument shares certain features with the

considerations that generate the Rowlands deadlock, it demands attention in its own right. At its heart is 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 traditionally 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 this volume b), if we take some putative case of extended cognition, we can always imagine a functionally equivalent system 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, because 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 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 an example of distributed (over brain, body and world) problem-solving that intuitively looks to be wildly unlikely candidate for a case of extended cognition, so unlikely in fact that any theory according to which the external parts of that system 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 entire distributed system described at step 1 must also count as 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 the Sprevak 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. Thus it does not entail that where the causal contribution to intelligent behaviour of certain in-the-head elements is functionally identical to that of certain noncognitive external elements, the former elements attain cognitive status purely in virtue of being intra-cranial. The orthodox version of the Martian intuition is also fully compatible with the parity principle. The path to the Sprevak dilemma is thus blocked, at step 2.

It is worth noting that the missing piece of the jigsaw here is some sort of locationally independent account of the cognitive that fixes the benchmark for parity (see section 5 above). Once such an account is part of our conceptual picture, there is no reason at all to think that any old process will count as cognitive, just because it has been rammed inside the head of a Martian. The resulting benchmark for parity does sterling theoretical work in weeding out unwanted interlopers into the domain of the cognitive, wherever they happen to be spatially located.[8]

**8. A Loose Ending**

In this paper I have argued that it is possible to defend the thesis of extended functionalism against some seemingly powerful objections. But perhaps this result, as encouraging as it is for the prospects of extended cognition, provides no grounds for a triumphant concluding flourish. It is common knowledge that functionalism *as a general theory of mind* faces some demoralizing philosophical challenges (for a nice review, see Levin 2008). Perhaps the more daunting of

these challenges are connected with phenomenal consciousness – the what it's-like-ness of experience. Who can forget evergreen thought experiments 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 such 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. Given the thought that phenomenal consciousness is central to mindedness, or at least to any mindedness interestingly similar to human mindedness, this looks like a serious limitation on any functionalist theory of mind, *including of course extended functionalism*. Extended functionalism inherits the disadvantages, as well as the advantages, of its parent theory.

A proper treatment of this issue must wait for another day. I simply want to bring the present discussion to a close by pointing out one thing. It is of course true that, to the extent that there exists a gap between functionalist explanation and an understanding of phenomenal consciousness, that gap is in force whether the realizing vehicles are wholly neural, a combination of neural and non-neural bodily factors, or an extended matrix of elements in the brain, the non-neural body, and the beyond-the-skin environment. But now notice that *it's the functional basis of the explanation that causes the alleged difficulty here, not where the realizing elements happen to be spatially located*. So although functionalism may indeed struggle in the face of phenomenal consciousness, extending one's functionalism certainly doesn't make things worse than they already were. When the topic at hand is the perplexing and recalcitrant question of how to account for phenomenal consciousness naturalistically, not making things worse is perhaps the best for which one can hope.

**Acknowledgments**

This paper was prepared thanks to support from the AHRC, under its Research Leave scheme, as part of project AH/F002963/1. Many thanks to Andy Clark and Peter Sullivan for useful discussions, and to Mark Rowlands and Mark Sprevak for permission to include quotations from unpublished manuscripts.

**References**


Adams, F. and K. Aizawa. 2008. *The Bounds of Cognition*. Malden, MA and Oxford: Blackwell.

Adams, F. and K. Aizawa. This volume. Defending the bounds of cognition.

Block, N. 1980. Troubles with functionalism. In C. W. Savage ed., *Minnesota Studies in the Philosophy of Science, Vol. IX*. Minneapolis, MN: University of Minnesota Press.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Churchland, P. M., 2005. Functionalism at forty: a critical retrospective. *Journal of Philosophy* 102(1), 33-50.

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

Clark, A. 1997. *Being There: Putting Brain, Body, And World Together Again*, Cambridge, Mass.: MIT Press.

Clark, A. 1999. Microfunctionalism: connectionism and the scientific explanation of mental states. English version of a paper that appears in German translation in T. Metzinger (ed.) *Das Leib-Seele-Problem in der Zweiten Helfte des 20 Jahrhunderts*. Frankfurt am Main: Suhrkamp. English version available on line at: http://www.philosophy.ed.ac.uk/staff/clark/pubs/microfx.pdf (accessed 16 July 2008)

Clark, A. 2005. Intrinsic content, active memory and the extended mind. *Analysis* 65 (1): 1-11.

Clark A. 2007. Curing cognitive hiccups: A defense of the extended mind. *Journal of Philosophy 104*, 163-192.

Clark, A. 2008. Pressing the flesh: A tension in the study of the embodied, embedded mind? *Philosophy and Phenomenological Research*. 76 (1): 37-59.

Clark, A. This volume a. Coupling, constitution and the cognitive kind: A reply to Adams and Aizawa.

Clark, A. This volume b. Memento's revenge: The extended mind, extended.

Clark, A. Forthcoming. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.

Clark, A. and D. Chalmers, D. 1998. The extended mind. *Analysis*, 58 (1): 7-19.

Doolittle, R. F. 1994. Convergent evolution: the need to be explicit. *Trends in Biochemical Sciences*, 19: 15-18.

Hurley, S. L. 1998. *Consciousness in Action*. Cambridge, Mass.: Harvard University Press.

Levin, J. 2008. Functionalism. *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), E. N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/fall2008/entries/functionalism/>. Accessed 19 September 2008.

Menary, R. 2007. *Cognitive Integration: Mind and Cognition Unbounded*. Basingstoke: Palgrave Macmillan.

Noë, A. 2004. *Action in Perception*. Cambridge, Mass.: MIT Press.

Putnam, H., 1967. Psychological predicates, in *Art, Mind and Religion*, eds. W. H. Capitan & D. D. Merrill. Pittsburgh: University of Pittsburgh Press.

Rowlands, M. 1999. *The Body in Mind*. Cambridge: Cambridge University Press.

Rowlands, M. 2003. *Externalism: Putting Mind and World Back Together Again*. Chesham, Bucks.: Acumen

Rowlands, M. Manuscript. Perception: from extended mind to embodied phenomenology. Unpublished paper given at a symposium on Embodied Perception held at the Spring 2008 meeting of the Pacific Division of the American Philosophical Association, March 18th-23rd 2008, Pasadena, California.

Rumelhart, D.E., Smolensky, P., McClelland, J.L. and Hinton, G. 1986. Schemata and sequential thought processes in PDP models. In *Parallel Distributed Processing: Explorations In The Microstructure Of Cognition, Vol. 2: Psychological And Biological* Models, J.L. McClelland and D. Rumelhart eds. Cambridge, Mass.: MIT Press, 7-57.

Rupert, R., 2004. Challenges to the hypothesis of extended cognition. *Journal of Philosophy* 101 (8): 389-428.

Smolensky, P. 1988. On the proper treatment of connectionism. *Behavioural and Brain Sciences*, 11: 1-74.

Sprevak, M. Manuscript. Extended Cognition and Functionalism. Draft paper available online at http://people.pwf.cam.ac.uk/mds26/files/Sprevak---Extended%20Cognition.pdf. Accessed 9 September, 2008.

Sutton, J. 2006. Distributed cognition: Domains and dimensions. *Pragmatics and Cognition*, 14 (2): 235-247.

Thompson, E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, Mass.: Harvard University Press.

Varela, F. J., Thompson, E. and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, Mass.: MIT Press.

Wheeler, M. 2005. *Reconstructing the Cognitive World: the Next Step*. Cambridge, Mass.: MIT Press.

Wheeler, M. Forthcoming. Minds, things, and materiality. In *The Cognitive Life of Things: Recasting the Boundaries of the Mind*, C. Renfrew and L. Malafouris eds. Cambridge: McDonald Institute for Archaeological Research Publications.

Wheeler, M. Manuscript. Draft chapters from an 'in preparation' book, *Extended X: Recarving the Biological and Cognitive Joints of Nature*, available on line (as of 11 September 2008) at http://www.philosophy.stir.ac.uk/staff/m-wheeler/ExtendedX.php

Wheeler, M. and Clark, A. Forthcoming. Culture, embodiment and genes: Unravelling the triple helix. *Philosophical Transactions of the Royal Society, series B*.

Wilson, R. A. 2004. Boundaries of the Mind: the Individual in the Fragile Sciences. Cambridge: Cambridge University Press.

---

[1] What I am calling the extended cognition hypothesis (ExC) trades under a number of different names, including the extended mind hypothesis (Clark and Chalmers 1998), active externalism (Clark and Chalmers 1998), vehicle externalism (Hurley 1998; Rowlands 2003), environmentalism (Rowlands 1999), and locational externalism (Wilson 2004).

[2] I am sensitive to the fact that the introduction of readily available electronic calculators and related software applications threatens to render this empirical observation about the use of pen and paper false. However, no one need worry about that here, since it is arguable that, with minor local variations to reflect how the different items of equipment are used, the points I go on to make in the main text apply straightforwardly to our relationship with the newer kind of technology.

[3] For my own contribution to this process, see Wheeler 2005. See also, among many many others, Varela et al. 1991; Clark 1997; Nöe 2004; Thompson 2007.

[4] I have set things up by treating ExC as a kind of radicalization of the embodied-embedded view (cf. Wheeler and Clark forthcoming). This is good enough for present purposes, but, as suggested in section 5 below, the relationship between the two sets of positions is a complex issue that has yet to be explored fully in the literature. For discussion of a number of apparent tensions between (certain versions of) the embodied-embedded view and ExC, see Clark 2008, forthcoming; Wheeler forthcoming.

[5] Of course, not all ExC theorists think that extended cognition should be justified by way of the parity principle. Arguments in support of ExC that don't exploit (and sometimes explicitly shun) the parity principle, are developed and defended by, e.g., Rowlands 1999, Sutton 2006, Menary 2007.

[6] Results from other psychological experiments on memory have been used in a similar way. For example, Rupert (2004) also appeals also to negative transfer interference effects (data which indicate that past learning interferes with the learning and recall of new paired associations), while Adams and Aizawa (2008) appeal to recency and primacy effects (data which indicate that we are better at recalling the elements at the beginning and end of a list than we are at recalling the elements in the middle). In both cases the claim is that extended systems will fail to exhibit the highlighted effect (or will fail to do so in the right way) and so are different in explanatory kind to the familiar human internal systems studied by cognitive psychologists.

[7] Equal treatment interpretations of parity based on (what I am calling) locationally uncommitted accounts of the cognitive are defended by Clark (2007, forthcoming) and Wheeler (forthcoming, manuscript). Within the scope of this general approach, there is a further and crucial question concerning how to unpack the key notion of a locationally uncommitted account of the cognitive. Clark (this volume b, forthcoming) suggests that the domain of the cognitive should be determined by our intuitive folk-judgments of what counts as cognitive. His supporting argument is (roughly) that our intuitive understanding of the cognitive is essentially locationally uncommitted, while the range of mechanisms identified by cognitive science is in truth too much of a motley to be a scientific kind, and so will thwart any attempt to provide a scientifically driven, theory-loaded account of the cognitive – locationally uncommitted or otherwise. I disagree with this assessment. I hold out for a locationally uncommitted account of the cognitive that is scientifically driven and theory-loaded, on the grounds (roughly) that our intuitive picture of the cognitive has a deep-seated inner bias, while Clark's argument for the claim that there is a fundamental mechanistic disunity in cognitive science is far from compelling (Wheeler, manuscript).

[8.] In effect, I have argued that the Mayan calendar program may be denied cognitive status, even when it is located inside a Martian head. This allows us to preserve the intuition that the Mayan desktop calendar program *as described* is the sort of element that ought to be excluded from the domain of the cognitive, an intuition with which of course Sprevak agrees. However, it is interesting to note just how sensitive our judgments are to the way the scenario is set up. For example, let's say we begin not, as Sprevak does, with the desktop program, but 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. This appears to be at odds with the conclusion drawn previously. Yet it seems all we have done is reverse the order in which the cases are considered. What is going on?

When we begin our reflections on the issues, as Sprevak does, by focusing on an example of a desktop program, our natural tendency is to think of an isolated and easily removable software application, sitting on a machine that sometimes achieves fancy feats of text-editing, graphics, and information storage, 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 cognitive, even if it were to be transported inside a Martian head. On the other hand, 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 a functionally integrated element in this kind of economy, then it may seem far less crazy to conclude that it could 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 we develop the basic idea, the resulting image is a long way from the one suggested by the scenario as described by Sprevak. In other words, the apparently fickle nature of our intuitions may be explained in terms of subtle changes to the details of the hypothetical example, changes that have been surreptitiously introduced by the variation in set-up.