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NB: Please do not quote or cite without permission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 3 Sameness and Difference

3.1 The Parity Principle

Under what conditions would it be plausible to say that cognitive traits may or do extend beyond the skin and into the environment? To launch our attempt to answer this question, here is a much-quoted passage from Clark and Chalmers.

> 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 these cases, the organism is linked with an external entity in a two-way interaction, creating a *coupled system* that can be seen as a cognitive system in its own right. All the components in the system play an active causal role, and they jointly govern behavior in the same sort of way that cognition usually does. If we remove the external component the system's behavioral competence will drop, just as it would if we removed part of its brain. Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head. (Clark and Chalmers 1998, pp.8-9)

This passage introduces what, in the ExC literature, has come to be known as *the parity principle*. In broad terms the parity principle as expressed by Clark and

Chalmers 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 the cognitive.

Clark and Chalmers' (1998) own example of cognitive extension through parity is the near-legendary (in ExC circles) case of Inga and Otto. In this imaginary scenario, Inga is a psychologically normal individual who has committed to her purely organic memory the address of the New York Museum of Modern Art (MOMA). If someone asks her the location of MOMA, she deploys that memory to retrieve the information that the building is on 53rd Street. Otto, on the other hand, suffers from a mild form of Alzheimer's, but compensates for this by recording salient facts in a notebook that he carries with him constantly. If someone asks him the way to MOMA, he automatically and unhesitatingly pulls out the notebook and looks up the relevant fact, viz. that the museum is on 53rd Street. Clark and Chalmers claim that there is a functional equivalence between (i) the behaviour-governing causal role played by Otto's notebook, and (ii) the behaviour-governing causal role played by the part of Inga's brain that stores the same item of information as part of her purely organic memory. By the parity principle, then, Otto's memory turns out to be extended into the environment. Moreover, argue Clark and Chalmers, just as, prior to recalling the information in question, Inga has the non-occurrent dispositional belief that MOMA is on 53rd Street, so too does Otto, although while Inga's belief is realized in her head, Otto's is realized in the extended, notebook-including system. It goes without saying that a good deal of the argumentative weight here rests on the details that mandate the judgment of parity. More on these in the next two chapters. For now we are concerned with the overall structure of the parity argument.

Although Otto's notebook provides the focus for many discussions of the parity principle, the basic idea is already bubbling away just below the surface of Dennett's (1996) discussion of "old folks" whose brains display "increasing imperviousness to new bouts of learning" (p.138) and who are removed from their homes to hospital settings. The behaviour of such hospitalized individuals, when compared to how they behave in their familiar home environments, often falls apart. Why is this? Dennett argues that over the years their domestic environments have become designed to support their needs by encompassing "ultrafamiliar landmarks, triggers for habits, reminders of what to do, where to

find the food, how to get dressed, where the telephone is, and so forth" (p.138). So far this might look like an argument for a merely embodied-embedded understanding of cognition, because of the way in which the designed home environments of the 'old folks' in question support everyday intelligence. However, Dennett draws the more radical ExC conclusion when he writes that "[t]aking [these people] out of their homes is *literally separating them from large parts of their minds – potentially just as devastating a development as undergoing brain surgery*" (pp.138-9, my emphasis). So what mandates the claim of cognitive extension here? It seems to be the thought that disrupting these people's environmental embeddedness has the same breakdown consequences as would disrupting their brains. And that, of course, is tantamount to saying, with Clark and Chalmers, that if "we remove the external component the system's behavioral competence will drop, just as it would if we removed part of its brain". So what would explain this equivalence between breakdown profiles? One compelling answer would surely be to point to a parity that exists between the functional contributions to the generation of behaviour made by certain states and mechanisms located in the brains of people whose brains remain "open to new bouts of learning" and the functional contributions to the generation of behaviour made by the external props and scaffolds that pervade the old folks' designed environments.

A final example of the parity-based argument for ExC demonstrates a subtly different way of appealing to the idea. As a bonus, it also shows that although many of the most discussed cases of putative cognitive extension turn on the contributions of artefacts (pen and paper props to reasoning, notebook supported memory), this is by no means an essential aspect of the ExC picture. Like the previous examples, however, it concentrates on a case of inner deficit compensated for by cognitive extension. Such cases provide tempting illustrations of parity, because they so clearly distinguish the familiar wholly inner case from the extended alternative, by introducing instances in which the purely inner solution is unavailable. But of course cognitive extension by way of parity, if it is plausible at all, is not limited to such cases. The unimpaired brain is at least as ripe for participation in extended systems as is the impaired brain. With that point made, on to the final example.

Susan Hurley (1998, 2003, forthcoming) discusses the case of an acallosal patient (someone whose corpus callosum is congenitally absent). The two hemispheres of such a patient's brain are disconnected. So here we have a subagential neural organization that is at best only partially unified. Nevertheless, this way of wiring things up (or rather of not doing so) may help to enable an agential

consciousness that is fully unified (e.g. speech and action are fully integrated). How is this possible? Hurley argues persuasively that the necessary integration may be achieved by feedback from active embodied vision. For example, side to side head movements may succeed in distributing information across the hemispheres, by enabling each half of the brain to receive direct sensory inputs from objects that would otherwise appear in only one half of the visual field, and would thus, in acollosals, be available to only one half of the brain. What justifies the interpretation of this arrangement as a case of cognitive extension, as opposed to merely embodied-embedded cognition? Hurley (forthcoming, pp.22-3) argues explicitly that the parity principle is at work.

> [B]y the parity principle, partly external processes could also enable integrated cognition; these could rely on bodily movements that distribute or transfer information across the hemispheres. Access movements – automatic, habitual side-to-side movements of head or body – could give each hemisphere direct sensory inputs from an object that would otherwise appear in only one hemisphere's visual field. Cross-cuing by automatic facial expressions accessible to both sides could also function to transfer information across hemispheres… Such extended mechanisms of integration would depend on bodily activity and feedback rather than purely neural factors. If they functioned when needed, reliably and automatically, by parity they would illustrate extended cognition.

What is striking about this appeal to the parity principle is a feature of the argument that is not remarked upon explicitly by Hurley herself. The mandate for cognitive extension does not flow from the familiar pattern of an identified functional isomorphism between the respective causal contributions of some purely external element (e.g. a notebook) and some purely internal element (e.g. a neurally located information store). Rather, it flows from an identified functional isomorphism between the respective causal contributions of a system that is already explanatorily spread (a system of access movements and cross-cuing alongside a direct contribution from the external objects themselves) and a purely internal element (a system of neural integration via the corpus callosum). Put another way, Hurley's example demonstrates that ExC may be justified not only on by appeal to parity between external and internal contributions to some cognitive trait (the information-store for memory may be external or internal), but also by an appeal to parity between a causally distributed solution to some cognitive problem and a purely inner one.

3.2 The Tyranny of the Inner

Now that we have encountered some illustrative examples of the parity principle at work, it is time to investigate the conceptual core of the resulting case for ExC, as revealed by the following question: what are the benchmarks by which parity of causal contribution is to be judged? Here is a suggested answer: First we fix the benchmarks for what it is to count as a proper part of a cognitive trait by identifying all the details of the causal contribution made by the brain. Then we look to see if any external or extended elements meet those benchmarks. One might think that this way of proceeding is invited by Clark and Chalmers' canonical formulation of the parity principle, as reproduced at the beginning of this chapter. Recall that, in cases of parity, "[a]ll the components in the [coupled] system… *jointly govern behavior in the same sort of way that cognition usually does*" (my emphasis). So now how are we to fix the meaning of the term 'cognition' in the rather vague phrase "in the same sort of way that cognition usually does"? One clue is provided by the observation that we are dealing with cases in which the test for parity, and thus for cognitive status, is a "part of the world [functioning] as a process which, *were it done in the head*, we would have no hesitation in recognizing as part of the cognitive process". Given this, one might take Clark and Chalmers' suggestion to be that parity *with the inner* determines cognitive status. But if that is right, then it seems to be no more than a short inferential hop to the related claim that in-the-head factors set the benchmarks by which parity of causal contribution is to be judged.

As things develop we shall see that this direct appeal to the inner reflects neither Clark and Chalmers' understanding of the parity principle nor the right way to understand that principle (which gives away the fact that, in my view, what I take to be Clark and Chalmers' understanding of the parity principle is not the right way to understand that principle). For the moment, however, let's follow the tabled suggestion through. What is immediately obvious is that it opens the door to the following strategy for resisting ExC: we identify certain features of some cognitive trait as standardly (internally) conceived that are not shared (or not shared in the right way – see later) by any extended arrangement that might be thought to perform the same cognitive task, and we conclude that since the parity principle is not satisfied, ExC is false. As the following passage demonstrates, Rob Rupert uses this kind of argument as part of his memory-oriented critique of ExC:

> I argue that the external portions of extended "memory" states (processes) differ so greatly from internal memories (the process of

> remembering) that they should be treated as distinct kinds; this quells any temptation to argue for [ExC] from brute analogy (*viz. extended cognitive states are like wholly internal ones; therefore, they are of the same explanatory cognitive kind; therefore there are extended cognitive states*). (Rupert 2004, p.407, my emphasis)

To substantiate this claim, 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 concludes that the effect is an accidental or optional feature, rather than an essential or definitional dimension, of the memory system.

Results from other psychological experiments on memory have been used in a similar way. For example, Rupert (2004) himself appeals also to negative transfer interference effects (data which indicate that past learning interferes with the learning and recall of new paired associations), while Fred Adams and Ken 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 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, which do. And, ironically, it is not only the critics of ExC who argue in this general way. ExC theorists in the complementarity-integrationist camp (more on whom later) sometimes reject parity-based arguments on exactly these grounds. Thus, commenting on Clark

and Chalmers' proposed functional isomorphism between Otto and Inga, Menary (2007, p.59) writes:

> Only at the grossest level of functional description can this be said to be true. Otto and his notebook do not really function in the same kind of way that Inga does when she has immediate recall from biological memory. There are genuine and important differences in the way that memories are stored internally and externally and these differences matter to how the memories are processed. John Sutton has pointed out that biological memories stored in neural networks are open to effects such as blending and interference (see Sutton [forthcoming] for discussion). The vehicles in Otto's notebook, by contrast, are static and do no work in their dispositional form (Sutton [forthcoming]).

Of course, in the hands of integrationists such as Menary, the final conclusion drawn is not that ExC is false, but rather that ExC cannot be established by way of the parity principle, because that way of arguing for ExC invites and succumbs to Rupert-style criticisms. Nevertheless, the driving observation remains the alleged lack of parity.

In order to see how the fan of parity-driven ExC might respond here, we need to understand the basic character of the criticism. This is one place where our tripartite analysis of ExC into general modal, human modal and human non-modal versions (see chapter 2) yields dividends. Given the central appeal to empirical data from psychology in the critical argument, it may look as if that argument is aimed at any parity-driven variety of the human non-modal version of ExC, the version which claims that at least some of the cognitive traits of an individual human thinker *are*, as a matter of fact, sometimes spatially located partly outside her skin. Indeed since (as noted previously) critics such as Adams, Aizawa and Rupert tend to draw only a two-way distinction between a modal and a non-modal version of ExC, and then proceed to endorse the modal version, it might seem that the argument *must* be aimed at the human non-modal version of the position.[1] However, armed with our three-way distinction, we can see that this isn't so. Rather, the target of the argument is any parity-driven variety of the human *modal* version, the version which claims that it is *possible* for at least some of the cognitive traits of an ordinary individual human thinker sometimes to be spatially located partly outside her skin. This is because the argument is that there are features of human memory that *cannot* be replicated in an extended memory system, or that where they can be replicated, their relationship with the

notion of human memory is somehow wrong (they are accidental rather than defining). So the phenomena to which the critics appeal (the generation effect, interference effects, recency and primacy effects etc.) are being wheeled in to contribute to a definition of *what counts as* human memory. Any system that does not exhibit them cannot be (is necessarily not) a form of human memory. So we are firmly in modal territory, although of course if the human modal version of ExC is false, so is the human non-modal version.

Now that we understand the line of attack, we can think about how an appropriate defence might be mounted. Let's begin by considering four replies that are suggestive but ultimately less than adequate, in order to clear a path to one that works.

The ExC theorist might try to exploit the fact that our parity gainsayers generalize from a conclusion about memory to a conclusion about all cognitive traits. The proposal here is that even if the Rupert-style argument undermines the idea of extended memory, that does not show that it will be equally successful when applied to other kinds of cognitive extension (extended reasoning, extended beliefs, and so on). There might be some mileage in such a response, given that the truth of the human modal version of ExC requires only that it is possible for *some* of the cognitive traits of an ordinary individual human thinker sometimes to be spatially located partly outside her skin. However, I am inclined to reject such a strategy. For one thing memory is, as Rupert (2004, p.23 manuscript) points out, an absolutely core cognitive trait. It is therefore plausible that what goes for memory goes for cognition in general. Furthermore, as we shall see, a better defence of ExC is available.

A second defensive strategy might complain that the critical argument under consideration simply begs the question against ExC, by assuming that what counts as cognitive should be fixed by the fine-grained profile of the inner. As it happens, the Rupert-style argument *does* beg the question against ExC, in precisely this way (Wheeler 2008). But, as Rowlands (manuscript) rightly points out, merely pointing that out is not sufficient to deflect that argument, since the question-begging would then travel in the opposite direction; that is, the proposed defence of ExC would be guilty of begging the question against the critical argument, by assuming that what counts as cognitive *should not* be fixed by the fine-grained profile of the inner. At best this produces a stalemate, and the critic of ExC might justifiably argue that in such circumstances there is a prima facie mandate for being conservative about one's chosen theory, and thus, given the received orthodoxy in cognitive science, for vehicle internalism.

A third defence of ExC here challenges the existence of the functional differences that putatively establish the crucial lack of parity. This might be done either by questioning the relevant empirical data from psychology or by exhibiting sufficient ingenuity so as to produce examples of extended systems that exhibit the effects in question in the right way. However, the former route is ill-advised in the case of widely observed effects and the latter has the air of ad hocery. In any case, as a general strategy, challenging the existence of every apparent functional difference between the purely inner and the extended versions of some target cognitive trait looks to be hopelessly misguided, since it seems beyond doubt that there will often be *some* such differences.

Regrouping again, the ExC theorist might claim that the precise way in which Clark and Chalmers allow the inner to set the benchmark for parity does not licence Rupert's view that the specification of that benchmark must be given in terms of the extant details of the human inner, as mapped out by conventional (i.e. skin-side) human cognitive psychology. The key thought here is that the regulating notion of the inner is not the one deployed by Rupert, but rather a counterfactual notion. And indeed, if we return to the canonical formulation of the parity principle, the counterfactual character of Clark and Chalmers' appeal to the inner leaps out at us: "[i]f, 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". It might seem that this reply neatly side-steps Rupert's argument. A problem is revealed, however, if we look in detail at the way in which Clark wields counterfactual cases in the heat of critique and counter-argument.

(Note: In what follows I take myself to be exploring a response that is suggested by certain passages in Clark when taken in isolation (see below), and to which Chalmers, as I understood him, appealed briefly in a recent lecture.[2] As will become clear later, however, I don't think the strategy in question ultimately represents either Clark's or Chalmers' rounded-out understanding of the parity principle.)

Consider the claim that, by the lights of the parity principle, Otto and his notebook constitute an example of cognitive extension. One reply to this claim begins with the observation that Otto's notebook is a system of essentially passive encodings. Such a storage system is, so the critic argues, importantly unlike our brain-bound memory systems, which support the propagation of new

information and the relevant updating of associated existing information in an active, automatic and unconscious way. This, we are told, is a functional dissimilarity that matters to the status of the Otto-notebook system as cognitive. (For one version of this kind of criticism, see Dartnall 2004. Notice that Sutton and Menary make a similar point about memory in an attempt to undercut the parity-based version of ExC, as part of their case for the complementarity-integrationist approach; see above.) Part of Clark's (forthcoming b) response to the tabled criticism is to deploy the counterfactual version of the parity principle. First we suppose that somewhere, inside some imaginary agent, there exists a system that realizes the very same functionality as Otto's notebook, which of course includes being integrated with other systems so as to support behaviour-generating activity. Then we test our intuitions: would we, without hesitation, count that inner system as part of our hypothetical agent's cognitive architecture? To encourage our intuitions in the right direction, Clark suggests the following example. Imagine we discovered Martians whose internally located memory systems worked like this: bit-map images of printed text are created and internally stored, and then later retrieved in the form of bit-mapped signals sent to and interpreted by the Martian visual cortex. We would surely have no hesitation in counting the bit-map system as part of Martian cognition. And so it should go for Otto's notebook and human cognition.

Now, I do not wish to question the thought that the application of the parity principle may, at least sometimes, proceed partly by way of a consideration of counterfactual or hypothetical cases. Such cases may play a perfectly good heuristic role, and maybe such a role is all that Clark, in particular, often has in mind (see e.g. his talk of the parity principle as a "rule of thumb" and as an "informal test"; Clark forthcoming b, p.3 and p.19 respectively). More on this later. What seems abundantly clear right now, however, is that *merely* shifting the ground from actual innards to counterfactual innards, such that those counterfactual innards take over the task of setting the benchmarks for parity, does not do enough to secure parity-based ExC against the Rupert-style critique. Why? Because while that critique is levelled against the human modal version of ExC, all that, at best, can be established on the basis of simple analogies with counterfactual innards such as Clark's bit-mapped Martian memory is the general modal version of ExC, a thesis to which Rupert and company typically don't object.

Let me explain. The first thing to note is that, strictly speaking, there is no need for aliens to be part of the story at all. Staying with the bit-mapped memory example, and assuming for the sake of argument that brain-bound human

memory is indeed active in the way described above, we can simply think about what things would be like if the human brain had been wired up by evolution so as to operate like Clark's Martian brain, and then proceed with the defensive argument from there. This is still a perfectly good example of the counterfactual innards strategy, and one that's human-centred. Clark (2005, p.7) pursues this sort of strategy in his example of counterfactual human innards for face recognition that are passive, discrete and photographic rather than active, fluid and reconstructive. However, whether we are dealing in exotic aliens or modified human beings, the fact remains that by imagining innards as different from (what we are taking to be) ordinary human innards as the bit-mapped memory system must be, the counterfactual innards strategy falls short of securing the human modal version of ExC, which after all states that it is possible for at least some of the cognitive traits of an *ordinary individual human thinker* sometimes to be spatially located partly outside her skin. We have seen already (chapter 2) that Clark sometimes has something close to human modal version of ExC in mind. Adapting the words he uses on such occasions, if we are interested in the human modal version of ExC, then we should be reflecting on "imaginable circumstances… that [involve] *no giant leaps of technology or technique*" (Clark forthcoming a, p.3). Now, exactly what counts as a giant leap here is not made clear, but I take it that major rewirings that transform the basic encoding structure of the human brain are ruled out. So any analogy with an inner memory that is like the bit-map example, in that it requires the kind of transformation just identified, whether that memory is inside Martian heads or human heads, cannot, it seems, do enough to establish the human modal version of ExC. The benchmark set by such innards is just the wrong sort of benchmark.

A tempting immediate response here is to claim that we should find a principled way to restrict the class of admissible counterfactual variations in innards, such that it includes only those variations that are appropriately close to what we know about the extant human inner, thus resituating our analogical model in the domain of the ordinary human. We can ignore the arguably vague nature of this suggestion, because, of course, many putative examples of ExC, such as the Otto-notebook arrangement, realize systems that are, in some ways, functionally very different indeed to those realized by ordinary human innards. So the adoption of the proposed restriction would merely place the ExC theorist back in the firing line of the Rupert-style critique.

The overall character of the move that the parity-driven ExC theorist needs to make should by now be clear enough. If human cognitive extension is to be judged generally possible, on the basis of parity considerations, then the extant

fine-grained details of the familiar internal manifestations of cognitive traits must not be allowed to set the benchmarks for parity. That's because the ExC theorist needs to argue that while the kinds of functional differences identified by Rupert and company clearly exist, they ultimately fail to tell against the parity-driven case for cognitive extension. Of course, the development of such a position must not lose sight of the fact that some functional differences undoubtedly will be relevant here. 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. What we seem to need, then, is some kind of scientific or philosophical theory that tells us which functional differences are relevant to judgments of parity and which aren't. One might see the strategy of using imaginary cases of counterfactual innards as analogical models as a failed attempt to meet this requirement.

3.3 Rethinking Parity

At this point someone might complain that I have been moving too quickly. Indeed, the need for a theory-loaded method for determining parity-relevant functional differences is, in effect, rejected by Clark (forthcoming b), when he argues as follows. (The quoted passage concentrates on the functional profile of stored information in dispositional believing, but the message is clearly designed to generalize.)

> [J]ust what aspects of the functional poise of stored information are essential if the information is to count towards an individual's stock of dispositional beliefs, and what aspects merely mark contingent features of current, standard human belief systems? Chalmers and I [in (Clark and Chalmers 1998)] tend to favor a rather coarse notion of the required functional role in which all that matters is that the information be typically trusted and that it guide gross choice, reason and behavior in roughly the usual ways. To unpack this just a tiny bit further, we can say that it should guide behavior, reason and choice in ways that would not cause constant misunderstandings and upsets if the agent were somehow able to join with, or communicate with, a human community. I do not see how to make this requirement any clearer or stronger without undue anthropocentricity. But nor do I see how to further argue this case with anyone whose intuitions differ.

Talk of guiding "gross choice, reason and behavior in roughly the usual ways" recalls an idea that we have met previously in the joint work of Clark and Chalmers, namely that of certain processes "jointly [governing] behavior in the same sort of way that cognition usually does" (see above). What is clear now is that Clark wishes to resist anything other than a kind of informal or intuitive understanding of the functional roles that set the benchmarks for parity. Anything more theory-loaded (my gloss on "clearer or stronger") than that, he warns us, will initiate a dangerous collapse into anthropocentricity. Let's agree that anthropocentricity is a bad thing, at least in the philosophy of mind. After all, one wouldn't want to rule out the plots of Star Trek, Star Wars, Doctor Who, Torchwood and Ben 10, as a matter of in-principle philosophical argument. Alien minds, should there be such things, are unlikely to work exactly like ours (I doubt that Vulcan memory exhibits primacy and recency effects), but they would be bona fide minds nonetheless. Still, why should *any* theory-loaded benchmark for parity have anthropocentrism as a consequence? Of course, if one assumes that the fine-grained details of the human inner, as mapped out by psychological science, constitute the only basis for a *theory-loaded* benchmark for parity, then anthropocentrism would be the only game in town. But if that assumption is what sits behind Clark's view, then he owes us an argument for it, because, as I am about to explain, it certainly looks to be false.

Here is a schema for an alternative theory-loaded benchmark by which parity of causal contribution may be judged (see Wheeler 2008 for an earlier version of the same idea). This proposal does not go via the fine-grained details of the human inner, and so invites no imminent collapse into anthropocentrism. First we give a scientifically informed 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. If this proposal can be made to fly, we would have a way of resisting Rupert-style arguments against parity-driven ExC, without begging any questions against the opposition, and without turning our backs on cognitive theory as the principal source of the benchmark for parity. So let's see if we can get airborne.

Peter Sullivan (in discussion) has suggested that however plausible the present proposal may be as some sort of weapon for defending ExC, it is debatable whether it is rightly portrayed as a way of unpacking the parity principle. This is because the role played in the new story by the scientifically informed account of

what it is to be a proper part of a cognitive system threatens to remove the need for any appeal to parity. If a candidate extended solution meets the conditions for being cognitive, as determined by whatever locationally uncommitted account of the cognitive we ultimately adopt (more on which in the next two chapters), the argument for ExC is already complete and parity considerations are made redundant.[3]

The right response here (I think) is to point out that the strategy under consideration may be explicated using a perfectly reasonable notion of parity, just so long as 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. What is less transparent though is how this equal-treatment notion of parity relates to what Clark and Chalmers mean, or perhaps are committed to meaning, by the term 'parity', either in their seminal jointly authored piece (Clark and Chalmers 1998) or in the salient parts of their solo-authored writings (e.g. Clark 2007, 2008b, forthcoming a, b; Chalmers 2008). To explore this issue, we can begin by considering some recent remarks from Clark (2008b, chapter 6, manuscript p.6) in which he attempts to clarify just how Chalmers and he originally understood the parity principle. (A similar passage appears in Clark 2007, manuscript p.7.)

> The parity probe was thus meant to act as a kind of veil of metabolic ignorance, inviting us to ask what our attitude would be if currently external means of storage and transformation were, contrary to the presumed facts, found in biology. Thus understood, parity is not about the outer performing just like the (human-specific) inner. Rather, it is about equality of opportunity: avoiding a rush to judgment based on spatial location alone. The parity principle was meant to engage our rough sense of what we might intuitively judge to belong to the domain of cognition—rather than, say, that of digestion—but to do so without the pervasive distractions of skin and skull.

As far as I can tell, this is an attempt to reject the Rupert-style fix-according-to-the-inner approach to parity and to replace it with an equal-treatment approach, while refusing to endorse the idea that equal treatment should be judged against some theory-loaded benchmark for parity. Clark's rejection of the theory-based option here is revealed by the fact that, consistent with what we have seen from him previously, he appeals to intuitive folk judgments about what belongs to the domain of the cognitive in order to provide the independent standard required by the equal-treatment approach. Such intuitive judgments stand in contrast to the scientifically informed, theory-loaded account of the cognitive that I have recommended should play the crucial benchmarking role.

One reason why Clark shies away from the sort of strategy I favour may be traced to his view that the range of underlying mechanisms that we collect together using terms such as 'mind' and 'cognition' is in truth too much of a mixed bag, too fundamentally disunified, too much of a motley to be a scientific kind (see Clark forthcoming b, pp.36-9). Building on this, Clark might argue that such a motley is unlikely to reward any attempt to provide a systematic theory-loaded account of what it is to be part of the cognitive. Of course, to give up on the prospect of a theory-loaded account of the cognitive is not yet to give up on the idea of a science of cognition, although a rethinking of the fundamental nature of the latter endeavour would be in order. As Sutton, a self-confessed member of the motley crew, argues, to pursue a science of disunified extended cognition, "it won't always be enough for distributed cognition [including ExC] enthusiasts to talk of the ecological validity of multidisciplinary immersion in the idiosyncratic and messy reality of cognitive practices" (Sutton 2006, p.2, manuscript). Some sort of conceptual or methodological glue is needed. To meet this challenge Sutton himself advocates, and indeed begins to develop, a taxonomy of the various elements that contribute to ExC explanations (e.g. embodied skills, external cultural tools and symbol systems), coupled with an analysis of the multiple and varied dimensions by which the distribution in evidence may be characterized (e.g. degrees of durability, degrees of medium-dependence of carried information, degrees of context-dependence, and so on). By contrast, Clark himself depicts the post-motley cognitive science as "a science of varied, multiplex, interlocking and criss-crossing causal mechanisms, whose sole point of intersection may consist in their role in informing processes of conscious reflection and choice" (Clark forthcoming b, p.39). (To maintain consistency in Clark's view, the "processes of conscious reflection and choice" mentioned here must be identified as part of our intuitive grip on the cognitive.) To their credit, then, Clark and Sutton both offer alternative conceptions of

cognitive science that do without a systematic theory-loaded account of what it is to be part of the cognitive. It seems to me, however, that such alternatives are the preferred option only if the motley-based argument against any theory-loaded account goes through. What, then, of that argument?

The first thing to note is that the scale of underlying disunity in the vehicles of cognition that Clark intends to convey by his talk of a motley is strikingly extreme – so extreme in fact that there will not even be "a family resemblance (at the level of actual mechanism) to hold [those vehicles] together" (Clark forthcoming b, p.16). This is not mere rhetoric. Clark needs the disunity to be this radical, if it is to undermine the attempt to provide a systematic theory-loaded account of what it is to be part of the cognitive. Of course, even a mild degree of disunity would rule out the chance of finding some small set of non-disjunctive necessary and sufficient conditions for a trait to be cognitive. But all that shows is that one particular kind of account is unlikely to yield dividends. Admittedly, if we really did arrive at the view that the vehicles of cognition defy even a family resemblance story, then the game would be up, but, as we are about to see, it's genuinely hard to see how contemporary cognitive science delivers a mandate for such a verdict.

When Clark (2008b, manuscript, chapter 5, pp.18-19) describes the patchwork of mechanisms that he takes to provide evidence for the claim that the vehicles of cognition are a motley, he identifies a number of dimensions of difference culled from the orthodox, and thus inner-oriented, psychological and neuroscientific literature. This temporary focus on purely inner mechanisms makes sense in the context of Clark's framework, since his "suspicion" is that there is no reason to think that the space of extended mechanisms will be any more disunified than the space of purely inner mechanisms (Clark 2008b, manuscript, p.19). This suspicion strikes me as correct. But one implication is that if the existence of a motley is not established by what scientific evidence tells us about the character of the inner, prior to any consideration of extended cognitive traits, then there is no reason to think that adding an extended dimension to the picture will produce a different outcome. Here, then, is the evidence that Clark submits (see Clark's text for supporting references). The mind as conceived by orthodox cognitive science seems to contain slow, conscious, controlled processes that degrade rapidly as cognitive load increases and that allow conscious interruption. However, it also seems to contain fast, automatic, uncontrolled processes that do not degrade with load or allow conscious interruption. It may well contain look-up-tables as well as more complex combinatorial representational formats. It seems to contain motor representations as well as

non-motor representations, and these different kinds of representation may support different episodes of imagination (e.g. mentally rehearsing a golf swing versus imagining a sunset over the sea). It seems to contain object-based visual representations that are abstracted from the egocentric particulars of the related visual stimulation, plus action-relative visual representations underpinned by neuronal cells that respond to a stimulus only when that stimulus is the target of, say, a grasping arm action or a saccadic eye movement.

The existence of this range of states and mechanisms is, of course, interesting and important, but it is not enough to establish the claim that the vehicles of cognition are a motley, in the sense that Clark requires. To explain: despite the nature of the evidence that he himself produces, Clark needs the alleged disunity in the vehicles of cognition to run deeper than the presence of multiple differences in things like, speed, controllability, or accessibility to consciousness. Indeed, if we take differences in algorithmic organization and high-level knowledge representation to be indicators of disunity, then even classical AI as practiced during connectionism's wilderness years would count as a motley (see Boden 2006 for the historical ebbs and flows of connectionist AI). Compare, for example, the STRIPS means-end analysis planning algorithm with the AQ11 incremental learning algorithm, or production rules with semantic nets. Surely the co-presence of these different symbol-processing algorithms and language-like representational structures in the vehicles of cognition would be no evidence of the kind of disunity in which we are interested, especially when all of them are presumably flag-bearers of Newell and Simon's (1976) famous claim that a physical symbol system (roughly, a classical AI system) has the necessary and sufficient means for general intelligence. And issues such as differences in speed, controllability and accessibility to consciousness can certainly be worked in to the classical story about our cognitive architecture without disrupting the basic unity of that architecture. For example, classical natural language processing algorithms for analyzing syntax, when implemented in human brains, were presumably supposed to be fast and resistant to conscious control, whereas many expert systems were supposed to capture aspects of conscious and deliberate problem solving. Any disunity here is, then, superficial.

To see if we can generate the kind of disunity that Clark needs, let's add mainstream connectionism into the picture and suggest that the vehicles of cognition are a combination of classical and orthodox connectionist systems. Our hypothetical cognitive architecture now features neurally inspired distributed representations and global computation by patterns of spreading activation alongside classical elements. Nevertheless, the fact is that, at a fundamental level,

all these different mechanisms, classical and connectionist, share a series of common deep assumptions about how cognition works, for example that it's essentially a matter of representational states transformed according to computational processes. So any motley-like qualities of the architecture remain essentially superficial. (For much more on the shared aspects of classical and mainstream connectionist thought, see Wheeler 2005b, chapter 3).

Where next? Work on the ways in which dynamical systems can underpin cognitive activity (see e.g. Port and van Gelder 1995) may well encourage us to be rather bolder in the way we construct our picture. I have argued previously (Wheeler 2005b) that our cognitive architecture is a shifting arrangement of (i) noncomputational, nonrepresentational dynamical systems, (ii) noncomputational, representational dynamical systems, and (iii) computational, representational dynamical systems. Is this the motley that Clark needs? I don't think it can be, since there is clearly a fairly straightforward family resemblance structure to this space of mechanisms: (i) and (ii) are noncomputational in character, (ii) and (iii) are representational.[4] In the end, then, contemporary cognitive science seems to provide no compelling evidence for the existence of a radical or fundamental disunity in the vehicles of cognition. And if there is no such disunity, there is no pressure on us from this quarter to be sceptical about the prospects for a scientifically informed, theory-loaded, locationally uncommitted account of what it is to be part of the domain of the cognitive.

The credentials of the theory-loaded account are enhanced further once we reflect for a moment on just what our pre-theoretical intuitions tell us about the domain of the cognitive. For as far as I can see, our contemporary pre-theoretical understanding of that domain includes a presumption of the within-the-skin internality of cognition that makes that understanding an unlikely source for the kind of locationally uncommitted benchmark for parity that ExC needs. I take it that the resistance that ExC typically elicits from the uninitiated is good prima facie evidence for the existence of such a presumption. And, in his early work on ExC, Clark conceded the possibility that the conclusion that Otto's dispositional belief about the location of MOMA is partly realized in his notebook goes beyond standard usage of the term 'belief' (see Clark and Chalmers 1998, p.14). More recently, however, Clark has claimed that any presumption of within-the-skin internality here is a result of the fact that "we are already in the grip of a form of theoretically loaded neurocentrism" (2008b, manuscript, chapter 5, p.35) and that the "folk grip on mind and mental states… is surprisingly liberal when it comes to just about everything concerning machinery, location, and architecture" (ibid. p.36). Of course, as Clark himself has persuasively argued,

various views from philosophy and science may, over time, become incorporated into what counts as our folk or pre-theoretical understanding of cognition (the Freudian unconscious would be one example; Clark 1989). So it is at least possible that what began as a theoretically loaded neurocentrism has now become an integrated *part of* our folk understanding of the domain of the cognitive, presumably through its capacity to give modern currency to a phenomenal sense of internality that preexisted contemporary scientific views of the brain as the seat of cognition. What this indicates is that the argumentative weight in Clark's position is carried not by the allegation of theoretically loaded neurocentrism but rather by whatever positive case can be built for the liberality of our folk understanding of where the mental might be located. To develop such a case, Clark (2008b, manuscript, chapter 5, p.36) appeals to an analysis of mental content and external representational media due to Houghton (1997).

Houghton's paper is remarkable in the way it anticipates a range of arguments and themes that have since come to be identified with ExC, including the parity principle itself (see e.g. Houghton 1997, p.170). That said, it is worth noting (because it will be important in a moment) that, right at the start of the paper, Houghton identifies his opponent as being the advocate of the view that mental states are "inner states in the strong sense that what mental states one is in depend solely on what obtains, or is going on, inside one's body, or, more particularly, inside one's head" (ibid p.159). Now, of course, as the distinction between content externalism and ExC (see chapter 2) shows, to claim that mental states *depend* partly on external factors is not yet to claim that mental states are partly *constituted* by external factors. So, to defeat his official opposition, Houghton need establish only content externalism. Nevertheless, his full-strength ExC convictions are clearly evident in passages such as the following. "The question of… whether something is part of a cognitive system, an accessory to it or part of its environment, is not one to be answered in the abstract. It is to encumber a future science of cognition with a wholly unreasonable burden if it is expected, in advance and independently of particular lines of enquiry, to decide whether certain extra-cerebral aids are part of our cognitive systems or not." (ibid p.172).

Within the details of this general foreshadowing of what we now recognize to be ExC, a distinctive aspect of Houghton's argument is that it finds support for cognitive extension not in the now-familiar examples of inner malfunction compensated for by distributed systems and solutions, but rather in examples of ordinary, everyday folk practices of attributing intentional states to agents who use external media to represent the world. For instance, Houghton argues that

the content of an architect's intentions regarding the detailed structural design of a future building need not, typically will not, and indeed sometimes could not, be fully internalized by the architect. Rather, it's the architect's drawings that give her intentions detailed content, by fixing that content even though her brain does not. On the basis of such examples, Houghton concludes that "it is undeniable that we commonly credit people with intentional states whose contents they themselves never fully internalize" (Houghton 1997, p.166). And later: to "deny that these are cases of genuine intentional content would be to use the notion of content in some way as yet to be explained which is clearly at odds with our ordinary attributions of intentional states" (ibid p.166).

Unfortunately there is a problem with Houghton's argument, one that reflects that apparent conflation of an externalism based on dependence with one based on constitution through realization. For while the practices of intentional-state attribution as embraced by the folk may well display an implicit endorsement of content externalism, it is far less clear that they display a similar endorsement of *vehicle* externalism, at least in anything like an unambiguous or consistent way. Thus the way in which the folk naturally attribute intentional content to an architect may well reflect the view that the content of an architect's drawings fixes the detailed contents of her plans and intentions regarding the nature of the building. And recognizing that fact has important philosophical consequences. But if we deliberately tried to prise apart our two forms of externalism, and so asked the folk a specifically vehicle-targeting question – for example, 'Where in space are the relevant cognitive states of the architect realized?' – I have no doubt that we would receive an internalistic, skin-side answer. Moreover, it seems, the practices of the folk reflect this answer. If an environmental protester had stolen the plans of Heathrow Terminal 5, the folk would mostly likely have been interested, and either supportive of the act or outraged by it, depending on what other beliefs were in play. But presumably none of these attitudes would be held because the folk were considering the whereabouts of part of Richard Rogers' mind. A plausible explanation for this pattern is that our folk grip on the cognitive involves a presumption of the within-the-skin internality of cognition.

Notice that I am not denying that Houghton's architect provides a case of genuine intentional content, the attribution of which would be natural given our ordinary practices. But this may be treated as an issue of content-individuation. Moreover, I am not disputing the ExC-style conclusion that the realizing vehicles of the architect's intentional states *are in fact* extended over brain, body and world. That may or may not be true. My point is that if this is a case of cognitive extension, then the fact that it is so suggests a theory-loaded revision to our

intuitive understanding of the domain of the cognitive, one that removes the presumption of the within-the-skin internality of cognition. This in turn implies that, pace Clark, our intuitive understanding of the cognitive is not apt to provide the locationally uncommitted benchmark for parity for which we are searching. That benchmark will need to be a theory-loaded construction.

Where are we? It seems to me that serious doubt has been cast on Clark's claim that the vehicles of cognition are characterized by a fundamental disunity. That was the claim that was supposed to motivate us to look to our intuitive judgments to provide a locationally uncommitted account of the cognitive as a benchmark for parity. And we have since discovered that, in any case, the presumption of the internality of cognition embedded in our folk understanding of the cognitive means that we have no good reason to think that our intuitive judgments about the domain of the cognitive are up to the task of generating such an account. Given his rejection of any theory-loaded alternative here, this leaves Clark without a locationally uncommitted account of the cognitive, and thus without the kind of benchmark for parity that would support an equal-treatment approach resistant to Rupert-style objections. So, this suggests that the application of the parity principle in support of the human modal version of ExC does ultimately require a benchmark for parity specified in terms of a theory-loaded, locationally uncommitted account of what it is to be part of the domain of the cognitive.

The same conclusion may be reached by a mildly different route. The line of argument that we have been following recently was prompted by the worry that by switching attention to the task of giving an appropriate account of the cognitive, we were kissing goodbye to the parity principle altogether. In contrast, I have suggested that we are rethinking rather than rejecting the parity principle, by virtue of the fact that we are now appealing to an equal-treatment notion of parity rather than an analogy-with-the-inner based one. Within this reformulated framework, we can make perfectly good heuristic appeals to concrete results from conventional skin-side human psychology and to imaginary cases of counterfactual innards, in order to probe various functional similarities and differences that might exist between a range of extended and non-extended solutions, just so long as we remember that when it comes to deciding where cognition is located, the all-important benchmark for parity is determined by some theory-loaded, locationally uncommitted account of the cognitive. I noted earlier that Clark sometimes writes as if his appeal to counterfactual innards in the application of the parity principle is no more than a heuristic tool of the sort just envisaged (see especially Clark forthcoming b, pp.3-4). It seems to me,

however, that this way of thinking is available only from a position in which access has been secured to an independent and appropriate account of the cognitive, and this is access which, if my recent arguments are correct, Clark doesn't have. In the absence of an independent and appropriate account of the cognitive, and given that the benchmarks for parity will need to come from somewhere, it is tempting to think in terms of a direct analogical route from the aforementioned imaginary cases of counterfactual innards to the determination of what counts as part of the cognitive. But, as we saw earlier, this kind of strategy ultimately leaves the ExC theorist vulnerable to Rupert-style arguments that exploit the functional disparities between wholly inner and extended cognitive solutions to undermine the parity-driven case for the human modal version of ExC. Once again, it seems that what is needed by ExC is a benchmark for parity specified in terms of a theory-loaded, locationally uncommitted account of the cognitive.

3.4 What Psychologists Care About

Naturally I wouldn't expect the positive position carved out during the previous two sections to be greeted with equanimity by the front-line critics of ExC, otherwise known (thanks to Mark Rowlands) as the *friendly hostiles*.[5] And indeed, Rupert (2004) and Adams and Aizawa (2008) present closely related arguments that might well be deployed in an attempt to resist the sort of position I have advocated. As we have seen, in a move that ought to be disputed by the ExC theorist, the friendly hostiles find their benchmarks for parity in the extant details of the inner. The new thought to be added in to the mix now is that this appeal to the inner is founded not on some pro-inner prejudice or some unwarranted conservatism, but rather on a healthy and entirely defensible respect for the methods and results of contemporary cognitive science, and in particular cognitive psychology. Thus Rupert (2004, pp.43-4 manuscript) 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." Similarly, after reflecting on what cognitive psychologists might plausibly find out by performing experiments on Inga and Otto, Adams and Aizawa (2008, pp.140-1) conclude that the functional differences between Inga's brain-bound memory and the Otto-notebook system are of a kind that cognitive psychology cares about, that what is going on in Inga's brain "answers at least very roughly

to what has normally been taken to be cognitive processing" and that this processing "is not very much like what is going on with Otto and his notebook". This result, we are told, "vindicate[s] orthodox cognitive psychology".

In general, then, the idea seems to be this: if we focus on the familiar experimental protocols adopted by cognitive psychologists, and we attend to the functional profiles that those psychologists care about, then we will (a) uncover a range of functional differences that distinguish purely inner solutions from extended ones, (b) provide a justification for the claim that the functional differences in question divide the cognitive from the non-cognitive, and (c) do so in such a way that the cognitive elements remain skin-side. Now, we know already that the ExC theorist simply must accept that there will be functional differences between purely inner and extended solutions. Indeed, we have spent a not-inconsiderable amount of time making conceptual room for that very result within the ExC framework. What this indicates is that point (a) will be shared ground between the two sides in the debate. But point (a), it seems to me, is as much as the ExC theorist needs to concede, because it is surely a mistake to think that just because a functional difference attracts the attention of cognitive psychologists, that difference must in some way play a role in marking off the cognitive from the non-cognitive.

A case of counterfactual innards, used as a heuristic device (see above), illustrates this point. 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 selective storage and context-sensitive retrieval of information. Nothing about this counterfactual case transports us into a universe populated by sentient aliens. It is also plausible, but I suppose open to dispute, that the removal of the generation effect would not require the human brain to be rewired extensively at some fundamental level, in order to implement a very different mode of information processing. Indeed, we don't assume that the brains of human beings who, say, recall Pi to many thousands of decimal places have been subject to root and branch neural rewiring. It seems, then, that the present example leaves us squarely in the realm of the human modal version of ExC. Moreover, I have no doubt at all that cognitive psychologists would find the functional difference between a normal human subject and our generation-effect-free subject interesting, and that they would use their well-understood experimental protocols to probe and explain it. But I cannot conceive of any cognitive psychologist concluding that the latter subject lacks the cognitive trait of memory, or even that she is utilizing some non-human kind of memory processing that cognitive psychology should properly ignore.[6] So why think that

exhibiting the generation effect is a defining dimension of (human) memory, rather than an accidental feature? And if that's right, then what is the justification (aside from our old friends, pro-inner prejudice and unwarranted conservatism) for refusing to apply the notion of (human) memory to an extended system with a similar profile to our generation-effect-free subject? To complete the case, and to drive home an earlier point, notice that this result is not restricted to memory. A similar argument could presumably 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.

The message here is not, of course, that no functional difference that ever interested a cognitive psychologist could ever be relevant to the issue of how to determine membership of the cognitive. The message, rather, is that working out whether or not a particular functional difference matters to this issue will not be decided by the fact that cognitive psychologists care about it or perhaps have investigated systems on one side of the divide it characterizes (e.g. systems that have the generation effect rather than lack it). For there will be functional differences that cognitive psychologists will want to investigate that nevertheless cannot be used to determine, of two traits separated by such a difference, that one is cognitive while the other is not, because the differences in question are differences *within* the domain of the cognitive. To repeat: from the mere fact that some specified functional difference happens to be of interest to cognitive psychologists, one cannot infer that that difference divides the cognitive from the non-cognitive.

This point against the critics is especially compelling if their key claim is understood as being that if a cognitive psychologist cares about a phenomenon such as the generation-effect, then displaying that phenomenon is a necessary condition for a system to be cognitive. However, we should pause to assess a less ambitious claim in the vicinity here. According to this alternative, although it is true, purely on the strength of what cognitive psychologists care about, that no particular phenomenon on its own, nor any particular group of phenomena considered jointly, may be used to specify a necessary condition for a trait to be cognitive, nevertheless, if enough functional differences between some target system and the objects of conventional cognitive psychology pile up, then we have good reason to exclude that system from the domain of the cognitive (cf. Adams and Aizawa 2008, p.140). But this claim can't be sustained either. For if the fact that conventional, inner-oriented, human-centred cognitive psychology has taken a professional interest in some phenomenon fails to place that phenomenon at the boundary between the cognitive and the non-cognitive (since

exhibiting or failing to exhibit that phenomenon may well locate a difference *within* the domain of the cognitive), then no matter how many phenomena of that kind one adds to one's list of absentees, one won't *thereby* have succeeding in marking out that boundary. On either interpretation, then, the pivotal inference in the critics' argument has not been established, which means that the anti-ExC conclusion isn't even on the horizon.

An initially more promising line of critical argument concentrates on the idea that the approach to securing ExC that I favour involves operating at such a stratospheric level of generality that it exists beyond the reach of decent psychological theorizing. For example, Rupert (2004) argues that any attempt to fix a generic notion of memory (a generic explanatory kind) that would subsume all the relevant internal and extended systems would need to be so devoid of detail (in order to subsume all the different profiles) that it would fail to earn its explanatory keep. Relatedly, Adams and Aizawa (2008, e.g. p.141) argue that it is hard to know what a theory of cognition would look like under conditions where that theory is supposed to be pitched at a level that is more general than that of conventional, wholly inner human cognition, while being general enough to cover an open-ended assortment of human-artefact ensembles, such that it succeeds in unifying all these different systems and processes. In tune with what I argued earlier in this chapter, I agree that a theory-laden account of what it is to be part of the cognitive that is general enough to allow for ExC, while nevertheless being imbued with appropriate explanatory bite, is something that the ExC theorist cannot do without (see also Rupert 2007, p.44, footnote 83, manuscript). But, unlike Adams and Aizawa, I see no reason to be sceptical about the possibility of generating such an account. Of course, such unbridled optimism will be vindicated fully only once we see what such an account might look like and what sort of pay-off we might get from adopting it. More on these issues in later chapters. In the meantime, however, we can nudge our intuitions in the right direction by considering again the case of the hypothetical subject whose inner processes of information storage and retrieval do not exhibit the generation effect. The fact that neither commonsense nor the interests of scientific psychology baulks at the idea that this subject's feats should count as cases of remembering gives us some reason to think that there must be a generic account of what memory is that is broad enough to cover generation-effect and non-generation-effect cases, and that has explanatory purchase. If this is right, then that account will presumably be apt to encompass, within the category of memory, extended mechanisms for information storage and retrieval that don't exhibit the generation effect.

To bring this particular line of argument to a close, it is worth considering briefly a different proposal for a benchmark-generating strategy that begins with the fine-grained details of the inner. According to this proposal, we should first work out the details of what the brain does, and then remove from our list of features any details that are inessential to that contribution *as cognitive*. That way we will be able to rule out the arbitrary exclusion of external elements and arrive at a viable set of criteria for what it is for something to count as a proper part of a cognitive system that does not rule out the very possibility of extended cognition. It seems, however, that any *decent* version of this response must collapse into a version of the strategy that I have been recommending, since in order for some detail of a causal contribution to be judged inessential to that contribution as cognitive, one must have access to an independent and locationally uncommitted account of the cognitive. The alternative, which would involve ruling out a detail purely on the grounds that it is not shared by some external element under consideration would, of course, beg the question against the opponent of ExC.

3.5 Sensing Trouble

It is time to consider a particularly awkward worry for any parity-driven argument for ExC. This worry turns on a seemingly obvious functional disparity that exists between purely inner cognitive solutions and most, if not all, putative cases of cognitive extension, a disparity which appears to be of a kind and an extent that significant doubt is cast on the cognitive status of the extended systems in question. Recall (yet again) the case of Inga and Otto. Otto's allegedly extended memory involves a stage in which he *writes the address of MOMA in his notebook*, and later stages in which he *looks at* that written information. It seems that nothing functionally similar to these events, the first a bodily action, the second a case of perception, happens when Inga stores and remembers the address of MOMA using her purely organic memory. Chalmers (2008, pp.2-3, manuscript) puts the worry like this. "It is natural to hold that perception is the interface where the world affects the mind, and that action is the interface where the mind affects the world. If so, it is tempting to hold that what precedes perception and what follows action is not truly mental." By this reasoning, Otto's notebook and the inscriptions in it do not count as realizers of part of his memory's information store, and so do not figure among the vehicles of cognition. And notice that the ExC-unfriendly, natural view that Chalmers describes is a product of our folk intuitions, which is another reason for suspecting that Clark's appeal to such intuitions, in order to delineate the

domain of the cognitive in an ExC-friendly manner, is in difficulty (cf. Chalmers 2008, p.3, manuscript).

One reply to the tabled objection (a generalized form of a reply to be found in Clark and Chalmers, 1998, p.16) is that the objection begs the question against ExC, because once we seriously entertain the idea that Otto's inner processes and his notebook constitute a single cognitive system, it is no longer true that Otto's writing movements and his visual access to the information so stored count as cases of action and perception. Instead they should be thought of as akin to the processes of information storage and retrieval that occur within the brain. As it stands this reply is, at best, radically incomplete. It depends on its advocate being able to give a non-question-begging (against the internalist) account of perception and action that (i) counts what happens at the periphery of the extended system as realizing the boundaries between mind and world described by Chalmers, and, crucially, (ii) fails to locate additional boundaries of the same kind at the edge of the sensory system and the limits of the organic body respectively, since these organic borders are no longer to be identified with such boundaries. Thus, to adapt a stock example, we need an account of perception and action that locates the relevant boundaries between mind and world at the limit of the blind man's cane, while also denying that the blind man perceives or acts on his cane. This looks to be a mighty big ask.

So what's the alternative? Chalmers himself opts for a kind of contextualist strategy in which whether or not the Otto-notebook system is a case of cognitive extension depends on which question we are asking (see Chalmers 2008, pp.4-5, manuscript). Concentrating on the perceptual access dimension of the case, here is how Chalmers' contextualism works. If our question is 'Why did Otto look in his notebook?', then it will be natural to say that he wanted to go to MOMA, did not know its location, but believed that its address was recorded in his notebook. Under these circumstances, Otto does not have the disputed extended belief and his memory is brain-bound. So his accessing of the notebook ought to be conceived not as an act of remembering, but as an act of perception. If, on the other hand, our question is 'Why did Otto take the subway to Fifth Avenue/53rd Street?', then Otto's interactions with his notebook become a background issue, and it will be natural to say that his beliefs and memory are extended. In this case, Otto's accessing of the notebook ought to be conceived not as an act of perception, but as an act of remembering.

It strikes me that there is a substantive difficulty with the way Chalmers walks us through the second of these scenarios. To see the problem, we need to begin by

pinpointing why is it that our answers to questions such as 'Why did Otto take the subway to Fifth Avenue/53rd Street?' have the effect of backgrounding Otto's perceptually guided interactions with his notebook. Presumably the crucial observation here is that, in answer to these sorts of questions, we are inclined to say "because Otto desired to go to MOMA and believed that MOMA was on 53rd Street". Call this the belief-desire explanation of Otto's general behaviour. It seems to provide a perfectly reasonable answer to the target question, and one that doesn't call for any mention of Otto's interactions with his notebook. Unfortunately, it doesn't follow from the backgrounding effect in question that the objection on the table is defeated. This becomes clear once we are more specific about what that objection actually says. The observation that Otto accesses the information in his notebook using his visual system is supposed to undermine the claims (i) that, *in advance of looking in his notebook*, Otto possessed the belief that MOMA is on 53rd Street and (ii) that the notebook is an information store that counts as part of Otto's long-term memory. But now the difficulty with Chalmers' analysis comes into view. The belief-desire explanation of Otto's general behaviour is certainly compatible with claims (i) and (ii), but it is also compatible with the denials of these claims. That is, it is compatible with the view that, *in advance of looking in the notebook*, Otto did *not* possess the belief that MOMA is on 53rd Street, and with the view that the notebook is an information store does not count as part of Otto's long-term memory. To see why, notice that it is agreed on all sides that, *once Otto has looked in his notebook*, he has the belief in question. The opponent of ExC understands this as a familiar case of internal belief formed on the basis of environmentally located information, and holds that it is this later-formed internal belief that figures in the belief-desire explanation of Otto's general behaviour. Moreover, it is also agreed on all sides that the information which figures in the content of that internal belief guides Otto's behaviour. However, the opponent of ExC denies that Otto may rightly be said to possess that piece of information prior to looking in his notebook and forming the internal belief just mentioned. So the notebook does not count as part of Otto's long-term memory. What all this suggests is that the fact that we background the matter of Otto's perceptual access to the information his notebook, when we explain his behaviour in a certain way, does not do enough to dispel the present objection to parity-driven ExC.

Fortunately, Chalmers-style contextualism is not the only game in town. What the ExC theorist needs to do is agree that there are genuine functional differences here (meaning that Otto's deployment of perception and action can't be assimilated to Inga's internal processing), while succeeding in showing that these functional differences have no negative implications for the status of the

extended system in question as cognitive. Although I don't know how to achieve this conclusively, I do think that serious doubt can be cast on the thought that the functional differences in question *must* have such negative implications. The key here is to realize that if perception and action are to erect a barrier to cognitive extension, then these phenomena have to be thought of in a particular way, that is as *interfaces between mind and world*. On the interfaces view, mind and thought are trapped inextricably *between* perception and action. This means first, that perception is theoretically disassociated from action, and second, that whatever elements lie on the worldly side of perception (as sources of inputs to the mind) or on the worldly side of action (as outputs of mind) cannot themselves be part of the realizing base of the cognitive. This seductive picture is explicit in the way Chalmers states the tabled objection (see above). It is also assumed by the traditional explanatory framework of orthodox cognitive science (as pointed out by Brooks 1991, Clark 1997, Haugeland 1995/1998, Hurley 1998, and Wheeler 2005b, among others) and has a long philosophical history that becomes tangibly manifest in Descartes' account of mind-world relations (Haugeland 1995/1998, Wheeler 2005b). But however natural or commonplace it may be, and whatever intellectual heritage it may have, the interfaces view is neither scientifically nor philosophically mandatory. Here is one way in which it might be challenged.

Researchers in the area of contemporary AI sometimes known as *situated robotics* favour the design and construction of complete robots that are capable of integrating perception and action in real time so as to generate fast and fluid adaptive behaviour. In pursuing this aim, situated roboticists shun the classical cognitive-scientific reliance on detailed internal world models, on the grounds that such structures are computationally expensive to build and keep up to date. In its place they put a design strategy according to which the robot regularly senses its environment in order to guide its actions. It is this specific behaviour-generating strategy that marks out a robot as *situated* (Brooks 1991).[7] Barbara Webb's cricket robot, described in chapter 1, is a paradigmatic example of the approach. We shall encounter further examples as our investigation progresses through future chapters.

[Note for those who have downloaded this draft chapter from the Web. I have not made chapter 1 available online, so here is a brief description of Webb's robot.

Consider the ability of the female cricket to find a mate by tracking a species-specific auditory advertisement produced by the male. According to Barbara Webb's robotic model of the female cricket's behaviour, here, roughly, is how the phonotaxis system works (for more details, see Webb, 1993; 1994; or the discussion in Wheeler, 2005). The basic anatomical

structure of the female cricket's peripheral auditory system is such that the amplitude of her ear-drum vibration will be higher on the side closer to a sound-source. Thus, if some received auditory signal is indeed from a conspecific male, all the female needs to do to reach him (all things being equal) is to continue to move in the direction indicated by the ear-drum with the higher amplitude response. So how is it that the female tracks only the correct stimulus? The answer lies in the activation profiles of two interneurons (one connected to each of the female cricket's ears) that mediate between ear-drum response and motor behaviour. The decay rates of these interneurons are tightly coupled with the specific temporal pattern of the male's song, such that signals with the wrong temporal pattern will simply fail to produce the right motor-effects.

Now, here is Webb's own explanation of why the mechanism just described is adaptively powerful: 'Like many other insects, the cricket has a simple and distinctive cue to find a mate, and consequently can have a sensory-motor mechanism that works for this cue and nothing else: there is no need to process sounds in general, provided this specific sound has the right motor effects. Indeed, it may be advantageous to have such specificity built in, because it implicitly provides 'recognition' of the correct signal through the failure of the system with any other signal' (Webb, 1993: p.1092). So the situated special-purpose adaptive coupling that constitutes the cricket phonotaxis mechanism works correctly only in the presence of the right, contextually relevant input.]

Building on the basic idea of situatedness as a principle of adaptive behaviour, one of the key lessons from situated robotics research is that much of the richness and flexibility of intelligence is down not to centrally located processes of reasoning and inference, but instead to integrated suites of special-purpose adaptive couplings that combine neural mechanisms (or their robotic equivalent), non-neural bodily factors, and environmental elements, as 'equal partners' in the behaviour-generating strategy. Unsurprisingly, then, the field of situated robotics is a rich storehouse of examples of embodied-embedded cognition and perhaps – if ExC can be made plausible, and if whatever conditions for parity we finally arrive at are met – of cognitive extension. For present purposes, however, the significance of the approach is that the principles and ideas just mentioned provide the platform for a refusal to conceptualize perception and action as interfaces between mind and world. As Rod Brooks (1991, p.173) puts it, one of the guiding principles of the approach is that: "There is no separation into perceptual system, central system [i.e. thought], and actuation system. Pieces of the network [the distributed robotic control system] may perform more than one of these functions. More importantly, there is intimate intertwining of aspects of all three of them." In other words, as John Haugeland (1995.1998, p.221) explains, what situated robotics tells us about

perceptually guided intelligent action is that the basic structure of such activity is one of "*interaction*, which is *simultaneously perceptive and active*, richly integrated in real time" (second emphasis mine). This dissevering of perception, thought and action in the context of situatedness undercuts the idea that perception and action are distinct interfaces between mind and world. It thereby relieves us of the motivation for holding that what precedes perception and what follows action *must* be non-cognitive. And that means that it is no longer obvious that the functional differences that clearly exist between Inga's organic memory and Otto's notebook-supported system of information retrieval, given the dependence on perception and action that characterizes the former but not the latter, *must* have negative implications for the parity-driven case for judging the latter to be a case of extended memory.

All in all, it seems to me that arguing for ExC by way of the parity principle, when that principle is understood in the right way of course, is a promising tactic in good conceptual order. The missing piece of the jigsaw is the much-lauded (by me) theory-loaded and locationally uncommitted account of what it is to be part of the domain of the cognitive. It's this that will provide us with our all-important benchmarks for parity. In the next chapter I shall launch a search for this missing item.

Notes

1. For the wrinkles in this interpretation of Adams and Aizawa's position, see chapter 2, note 2.

2. David Chalmers' Barwise Prize Lecture, The American Philosophical Association Pacific Division, Eighty-Second Annual Meeting, Hilton Pasadena, Pasadena, California, March 20, 2008.

3. It is not lost on me that once our attention is deflected away from considerations of parity, my plea for a scientifically informed and locationally uncommitted account of the cognitive is reminiscent of Adams and Aizawa's (2008) demand for a *mark of the cognitive*. This is an issue to which we shall return in chapter 4.

4. My analysis assumes that computational systems are a restricted subset of dynamical systems (see Wheeler 2005b, chapter 4, for the arguments). This is not a universally shared view (see e.g. van Gelder 1998), but even if we adjust for

this, and so make (iii) a case of computational, representational, *nondynamical* systems, the highlighted family resemblances will still apply.

5. At the *Extended Mind II* conference (University of Hertfordshire, July 2006), Mark Rowlands complained that the terms he introduces never catch on. I use his phrase 'the friendly hostiles' to right that heinous wrong.

6. At one point Adams and Aizawa (2008, p.140) argue that if we reflect on what cognitive psychology is likely to tell us about the differences between purely inner memory systems, such as Inga's, and related extended solutions, such as the Otto-notebook system, our conclusion should be that there is a "distinctly human kind of memory processing" which is exhibited by the former but not by the latter, and that this "warrants cognitive psychologists in pursuing an intracranial human cognitive psychology that is not based on mere prejudice". One natural reading of this claim is that, for Adams and Aizawa, it is possible to think of extended systems such as the Otto-notebook system as realizing a non-human kind of memory processing that cognitive psychology should properly ignore.

7. In contemporary philosophy of mind and cognitive science, the term 'situated' is often used to mean 'environmentally embedded'. In this book I shall use it more narrowly, to name the specific behaviour-generating strategy of regularly sensing the environment to guide action rather than using some detailed internal world-model. As the main text goes on to indicate, one might think of this strategy as underpinning a form of environmental embedding.