

Philosophical Studies Book Symposium on *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (Oxford University Press, NY, 2008)

Précis, and Author's Response.

Précis of *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (Oxford University Press, NY, 2008)

Andy Clark

Supersizing the Mind (henceforth, SSM) was a book with a double mission. The first mission was to display and discuss the rich and varied landscape of recent work in the area of (broadly speaking) embodied, environmentally embedded, cognitive science. To this end, I canvassed and organized a wide range of examples in which fine details of embodiment, of worldly action, and of worldly resources, could be seen to make diverse and unexpectedly deep contributions to human cognitive achievements. The second mission, building in many ways upon the first, was to pursue the more radical suggestion that is now known as the Hypothesis of Extended Cognition¹. This was the suggestion² that in some such cases, there is a sufficiently dense degree of inter-animation between the neural and the gross-bodily, or even between the organismic and the extra-organismic, for it to become ill-warranted and unproductive to reserve the label of 'cognitive processing' for the inner, neural or organismic contributions alone. Of course, the mere fact of dense inter-animation will not be enough: there may well be dense inter-animation between, say, the sailor and the sailboat, or between the

digestive tract and the brain, without either the sailor-sailboat or the brain-digestive tract system counting as an extended *cognitive* system. But where we find dense inter-animation *and* that inter-animation looks to be serving recognizably cognitive (for example, broadly speaking epistemic or knowledge-oriented) ends, then (assuming, see below, that we can also assign ‘ownership’ of the relevant states or processes to a distinct agent) then there is – or so I argued - no good reason to carve the mental cake according to merely metabolic joints: no good reason, that is, to think that all our cognitive processes need be found in the head, or even within the biological organism.

The first mission (that of displaying the broad shape of ‘embodied, embedded cognitive science’) is important, though it attracts – perhaps understandably - far less critical attention than the second. In pursuit of the first mission, SSM highlighted the notion of ‘information self-structuring’ (Lungarella and Sporns (2005)). The key idea here was that at multiple time scales, embodied agents structure their own information flows in ways that support richer forms of adaptive and cognitive success. Sometimes, such structuring happens on the short time scale of bodily movement. For example (Ballard et al (1997)) when we move our heads and eyes as we speak and reason. It also happens on intermediate time scales, for example when we scribble words on a page while attempting to plan just when we need to get the taxi to arrive at the bus station in time to catch the bus to the airport. And it happens on even longer time scales when we structure our persisting environments in ways that will serve future needs, as when we leave a yellow sticky note on the door reminding us to get milk, or post signs on the highways

directing motorists to their destinations. In all these cases we (either individually or collectively) structure our worlds and actions in ways that - usually productively - alter the flow of information arriving at the biological brain. It is these abilities of rampant and iterated 'cognitive niche construction' that have set us somewhat apart from the other animals with whom we share both a planet and a massive fundamental biological heritage.

SSM was also at pains to distance itself from those versions of embodied cognitive science that reject the use of notions of internal computation and/or internal representation in their explanations of human thought. It was at pains, too, to avoid accounts of thought and experience that (it was argued) give *too strong* a role to the idiosyncrasies of human embodiment and action. More positively, SSM suggested that we now possess the main tools needed to arrive at a mature science of the embodied mind. What is needed is a careful combination of computational, representational, and dynamical sensibilities: one that recognizes that bodily motions (as in the case of eye fixations, gesturing while speaking etc) may themselves be playing important information processing roles. Given this kind of 'extended functionalist' approach, we might even (see chapter 9 of SSM) one day hope to *quantify* (Lungarella et al (2005)) the contributions of embodied action to cognitive processing using various information theoretic measures.

Mission 1 phased directly into the rather more contentious (philosophically at least) mission 2: the defense of the claim that, under certain circumstances, the dense interplay between neural and extra-neural factors might be such as to warrant

talk of extended cognitive processes or even of an 'extended mind'. The argument presented comprised three key components. First, there was a general principle (the so-called Parity Principle, more on which below) which is best seen as a heuristic (a rough-and-ready tool) for identifying some plausible cases of cognitive extension. Second, there was a thought experiment (the case of Otto) meant to convince the reader that, under certain conditions, the coarse functional role of a bio-external encoding could be sufficiently similar to that of a persisting internal encoding as to mandate similar treatment, revealing the non-biological resource as part of the physical machinery underpinning some of an agent's genuine mental states. Third, there was an attempt to display the specific kinds of temporal and computational complexity that might further support the even-more-contentious claim that various relatively transient real-world systems may also be profitably analysed – while up-and-running, so to speak - as unified cognitive wholes, rather than by splitting them into a cognitive (biological, typically neural) component and various ('non-cognitive') sources of input, transformation, and storage.

Here is a super-brief sketch of each of these three components, starting with the rule of thumb now known as the Parity Principle. The parity principle, as introduced by Clark and Chalmers (1998, p.8) suggests that "If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process." The idea here was simply to invite the reader to judge various potential cognitive extensions behind a kind of

‘veil of metabolic ignorance’. A good way to do this is ask yourself, concerning some candidate cognitive process P, whether *if* you were to find P (or better, its functional equivalent) occurring inside the head of some alien organism, you would tend to class P as a cognitive process? If so, then the onus – it was claimed - is on the skeptic (the person who wants to deny that the functionally equivalent process P, when bio-externally realized and suitably coupled with the biological agent, is a cognitive process or forms part of the agent’s cognitive processing) to make her case.

Alongside this simple rule of thumb, Clark and Chalmers offered a thought experiment meant to show how, for the familiar case of an agent’s dispositional beliefs, a bio-external resource might indeed make the grade. The thought experiment concerned Otto, a person with mild memory impairments, who makes extensive use of a notebook to guide his behaviour. I won’t rehearse the Otto case yet again in this short précis, except to note that the crucial point here was that by treating the notebook encodings as directly partially determining Otto’s some of standing beliefs, we get to grips with the *very same coarse patterns* in Otto’s behaviour as we do, in more standard cases, by treating an agent’s neural states as such realizers. For example, we lock onto a coarse pattern characteristic of holding the standing belief that MOMA is on 53rd street by treating the inscriptions in Otto’s notebook as partially realizing some of *his* standing beliefs, just as we might do with a normal agent (‘Inga’) by treating some of her neural states as such realizers. Both Otto and Inga, we argued, should be treated as holding the standing belief (that is, as believing even before the moment of conscious recall) that MOMA is on 53rd Street. Unlike Inga,

Otto is not a fully normal agent, but courtesy of the extra-biological machinery, much more of his behaviour can be successfully subsumed under our familiar folk psychological kinds: kinds such as ‘believing that MOMA is on 53rd street’.

The third and final ingredient in Mission 2 was an attempt to plot, in rather more detail than before, something of the kinds of temporal and computational complexity that characterize the best real-world exemplars of extended cognition. The goal here was to highlight the potential complexity of the ‘dovetailing’ that might be achieved between plastic neural systems and highly practiced bio-external props and supports. An important upshot of such complexity, I argued (here pursuing ideas that I first encountered in Kirsh and Maglio (1994)) was that there need be no neat, persisting ‘agent-level’ bottleneck mediating the brain’s calls to some external resource. Instead, different neural sub-systems would have their own sub-personally mediated ways to ‘call’ or access the external resource, thus building the resource’s reliable presence so deep into the information-processing flow chart that it becomes visibly arbitrary and unhelpful to draw a single metabolically determined line dividing the truly ‘cognitive’ (i.e. participating in the cognitive processing of the agent) aspects of that processing from the rest.

SSM presented various bodies of research (from the use of deictic pointers in visual processing, to multiple timescale coupling in Tetris, to recent systematic work by Gray et al (2006)) on the cognitive control of interactive behaviour) meant to lend support to this vision of complex dovetailing. Such complex dovetailed wholes will genuinely reward, or so

SSM argued, understanding in terms of an extended functional organization: an extended functional organization relative to which the metabolically-determined inner-outer boundary is both analytically unhelpful and computationally far less significant than one might have pre-theoretically supposed.

None of this was meant to suggest that in such cases an extended perspective was mandatory, or (worse still) that in such cases the neural contribution to human cognitive success was somehow ‘less special’ than we might have previously imagined. Indeed, a large chunk of SSM is devoted to displaying the human brain as an organ supremely well equipped to thus recruit external structures and opportunities deep into its problem-solving routines. It is only once this important work has been achieved, SSM suggests, that a more egalitarian regime is enabled: one in which no special status accrues to the inner over the outer. In the end, then, the value of the extended perspective was said to depend on the fruitfulness of making a ‘mental flip’ that sees many of our genuinely cognitive unfoldings as running on machinery spread across brain, body, and world. It is always possible at that moment to flip back, and to see only the fine weave of inputs to, and outputs from, a complex inner machine. But to *confine* ourselves to that single perspective is to unreflectively privilege the inner and the biological in ways that, if SSM is on the right track, are both philosophically unmotivated and scientifically unsound.

Author's Replies

I was delighted, excited, and constructively challenged by this wonderfully engaging, illuminating, and diverse set of responses to *Supersizing the Mind* (SSM). I am immensely grateful to the three respondents (Wheeler, Hutchins, and Rupert) for the skill, energy, imagination, and patient good will with which they addressed the text. Rupert felt I went too far, Hutchins felt that I didn't go far enough, while Michael Wheeler, working from inside the extended mind camp (as it were), questioned a key step along the way. That full spread of responses gave me a sort of Goldilocks feeling, as if just possibly SSM was pitched about right!

Wheeler raises an important question concerning the structure of the argument in SSM, challenging my suggestion that friends of the extended mind can (and should) make do without first offering some kind of principle theoretical characterization of what *makes* a process a cognitive process. This is the debate concerning the need – or lack of one – for what Adams and Aizawa (2001) (2008) call a 'mark of the cognitive'. Rupert, continuing in his role as careful critic of the claim that cognitive processes extend into the non-biological realm, suggests that nothing in SSM serves to repudiate his own suggestion that we ought rather think (and think only) in terms of persisting integrated cognitive systems – systems which turn out to be nothing but the organismic wholes themselves - productively embedded in empowering, partially self-engineered, environments or niches. Hutchins, writing from his signature perspective within cognitive anthropology, worries that SSM, by stressing the pivotal role of the brain in the recruitment of external resources and in

the maintenance of resource-engaging cycles, actually gives too much away to a traditional internalist vision. I shall say a few words about each of these responses in turn.

Wheeler notes, correctly, that the intended reading of the Parity Principle (see Précis above, and discussion in SSM chapter 4) was a plea for equality of opportunity rather than equality of mechanistic contribution. The idea there was not that external stuff must work in much the same way as inner stuff if cognition is to depend on extended mechanisms. Rather it was to probe how we would treat the functional analogues of certain external contributions were they (appropriately) internally relocated. Unfortunately, both friends and foes of the extended mind have often misinterpreted the Parity Principle as indeed requiring some kind of functional equality between an existing internal cognitive mechanism and a putative external one. The mistake is understandable, since in the Otto thought experiment (again, see Précis and SSM chapter 4 and Appendix) Clark and Chalmers stress that the notebook entries govern Otto's behaviour in the same coarse-grained way as would the inner encodings of a normal agent. But all we meant by this is that *for most ordinary folk psychological purposes*, we lock on to many of the very same patterns in Otto's actual and counterfactual behaviour by treating the notebook entries as part of the mechanical supervenience base for his standing beliefs. In that restricted sense, and only in that restricted sense, are the two resources said to govern behaviors in similar enough ways. Importantly, this is something they can thus do despite a multitude of other more subtler differences in matters such as patterns of response to damage, possible sources of breakdown, and various

experimentally documented effects on memory and recall, such as recency and primacy effects etc. There is no need for a mark of the cognitive, I argued, because we already have an implicit (though probably totally unformalizable in words) grip on the kinds of *coarse-grained behavior patterns* that we take to be indicative of key mental states, such as the holding of a standing (dispositional) belief.

Wheeler is suspicious of this direct appeal to folk psychological intuition. Our basic folk intuitions, he points out, are about the way mental contents (beliefs etc) govern behaviour, and hence do not make easy or reliable contact with the kinds of questions at issue in debates concerning the extended mind. This is because the latter are really questions concerning the way that mental contents may or may not be legitimately ‘vehicled’ or realized in the material universe. Direct appeal to the folk intuitions, Wheeler claims, will either yield the wrong results (because the folk are basically internalists at heart), or will be of little value anyway, in debates concerning vehicles or realizers. Instead of appealing to the unregimented folk intuitions then, Wheeler suggests we should seek “a scientifically informed, theory-loaded, locationally uncommitted account of the cognitive”. We then apply this theory and see where the chips (the legitimate realizers of genuine cognitive processes) fall.

This is a splendid idea, but one that (it seems to me) is almost certainly doomed to failure. The reason it is doomed to failure is that the shape of any such scientific theory of legitimate vehicles will surely be determined, in large part, by what we take as central examples of real-world realizers of cognitive processes in the first place. The strategy of SSM was

thus to first try to shift the space of accepted exemplars of realizing systems by mobilizing our folk (and content-led, I agree) grip on the realm of the cognitive, but doing so behind the counterfactual ‘veil of metabolic ignorance’ that the Parity Principle was meant to provide.

Wheeler, however, offers a rather different reconstruction of SSM’s appeal to folk intuitions. That appeal, he suggests, must be rooted in my worry (e.g. SSM p.95) that even the inner vehicles of standardly accepted cognitive states and processes might themselves turn out to be an unruly motley, lacking even a useful kind of family resemblance. If that were so, Wheeler allows, the realizers would ipso facto resist regimentation into a principled and locationally uncommitted theory (the kind that might have then been deployed properly to adjudicate the debates concerning cognitive extension) since they would resist regimentation into a principled theory at all.

Wheeler’s counterfactual (if motley, then no principled scientific theory) is obviously correct. It is, after all, precisely the lack of a unifying scientific story that would justify the claim that we here confront only a motley. But Wheeler’s larger reconstruction makes the motley considerations argumentatively crucial in a way that I did not intend, and would not endorse. There is a clue to this in the text where I speak, as Wheeler himself notes, only of my ‘suspicion’ (SSM 95) that the inner goings-on will turn out to be such a radical motley. It would be a brave theorist indeed who left a pivotal argumentative move hostage to such uncertainty. I was not so brave, and this is fortunate since I no longer think the neural

realizers of cognitive states and processes are likely to form a motley after all.

In tentatively floating the idea of an inner motley, I relied, as Wheeler nicely shows, on an inadequate set of considerations. Roughly speaking, I pointed out a lot of apparent variety in the surface algorithmic forms of the neural underpinnings of abilities such as visually guided response, where (for example) categorization plausibly relies upon very different forms of encoding from fast, fluent visuomotor response (see Milner and Goodale (1995) (2006)). Similar points can be made regarding other domains and abilities. Nonetheless, as Wheeler notes, such (real) differences might nonetheless be built upon some kind of common underlying computational structure or form, for example the generic ‘rule and symbol’ forms of classical AI. The disunity might thus be superficial, masking a deeper unity that could still serve as the basis for a unifying scientifically informed theory.

But here’s the rub. Suppose that all our agreed cases of cognitive abilities do indeed succumb to a sufficiently unified theory of their *inner* (neural) underpinnings. Why suppose that that yields a “scientifically informed, theory-loaded, *locationally uncommitted* account of the cognitive” (my stress)? This is ‘locationally uncommitted’ only in a very weak sense viz that processes *just like that*, no matter where they are located, will count as cognitive processes. This actually sounds very much like the strategy recommended by classic opponents of cognitive extension such as Adams and Aizawa (2008). Given that brains are indeed pretty unusual bits of the physical universe, it seems overwhelmingly likely that what such a strategy will yield is an account of the cognitive heavily biased

towards its own origins, viz, the search for common *neural* threads underlying all forms of cognitive success. Such a story will, I shall now argue, leave all the key questions concerning cognitive extension unresolved.

To see this, let's 'plug in' an account that now strikes me as a very promising, and remarkably unified, model of the basic neural strategy underlying many varieties of intelligent response. I refer here to the emerging 'predictive coding' vision of the neural economy according to which (see e.g. Lee and Mumford (2003), Hohwy (2007), Friston (2005) (2010)) the basic work of the brain is to implement processes that correct errors in the prediction of input. In mammalian brains, such errors look to be corrected within a hierarchical cascade of cortical processing in which higher-level systems attempt to predict the inputs to lower level ones on the basis of their own emerging models of the causal structure of the world (i.e. the signal source). Errors in predicting lower level inputs cause the higher-level models to adapt so as to reduce the discrepancy. Operating over a plethora of linked higher-level models, the upshot is a brain that encodes (using different kinds of generative models for different purposes, thus recreating the superficial variety mentioned earlier) a rich body of information about the source of the signals that regularly perturb it. This model, it has recently been argued (Friston (2009) (2010)) provides a general, unifying account of the brain's capacities for learning, inference, and the control of plasticity, and provides a common framework in which to understand perception, action, and attention. This picture thus turns out to offer a breathtakingly comprehensive take on neural organization.

Suppose (just suppose) that this is the *right* model of our fundamental neural organization. We can (and should) still raise exactly the same questions concerning the potential role of the non-neural body and the extra-neural world in the construction of human thought and reason. One reason for doing so is that the mooted unifying account is indeed a unifying account (as one might expect) of neural – specifically cortical – microcircuitry. It is, moreover, microcircuitry we share with many other creatures who are notably (if sometimes adorably) less intelligent than ourselves. Perhaps then - and this is a theme taken up by Ed Hutchins in his insightful commentary – some of the key differences in animal mindfulness are not, or not all, fundamentally neural ones. This is, of course, exactly what the extended mind theorist would expect.

From within the neurally-unifying predictive coding framework, we can then ask about the potential role of (for example) gross bodily movements and long-term environmental structuring. The deep point of the predictive coding regime, according to e.g. Friston (2009) (2010), is that it allows brains like ours to minimize ‘informational surprise’ in their exchanges with the world. In perception, assimilating inputs to good generative models in the brain reduces informational surprise (prediction error). The more you do that, Friston argues, the more likely you are to survive and thrive. But that same overarching purpose can, as Friston himself notes, be served by bodily motion. By moving, we can actively select a less surprising set of inputs. The self-structuring of information flows stressed in SSM can thus be seen as a key device for minimizing informational surprise over various time-frames. After successful learning, we

mostly move so as to minimize surprise in the here-and-now, seeking the inputs we have come to expect. By contrast, as part of the process of learning, we sometimes move in ways that yield maximal information, so as to minimize surprise relative to a much longer time frame. The overarching goal of minimizing informational surprise can also be served (as Friston, personal communication, agrees) by the canny longer-term structuring of an environment, as when we write down our ideas while thinking, put signs on shops, paint arrows on country walks, etc. Our highly structured congenitally predictive brains are located in active bodies, in social networks, and in multiple forms and layers of technological scaffoldings. This leads to multiple, and multi-scale, cycles of information self-structuring, both as individuals and as a species. There is, in short, a kind of ‘family business’ here in which the brain, the body, and the self-structured (at multiple time-scales) environment all pitch in.

Given all that, should we just regard the neural kernel, the common neural mechanisms for the progressive reduction of prediction error, as limning the space of the cognitive? Or should we allow that genuinely cognitive processes can also become hybridized, so that *their* effective mechanisms include not just the neural elements but span brain, body, and world? This does not seem plausible in the case of the arrows painted on the country walk. But this may be because a raft of further necessary conditions is not met. In fact, a great deal of SSM (and work on the extended mind in general) is best seen as an investigation of such further conditions: conditions which must be met so as to ensure the *proper ownership* of some candidate extended process by a distinct cognitive agent

(for this argument, see SSM chapter 5). (Perhaps we ought rather to speak here of ‘proper inclusion within a distinct cognitive agent’ rather than ‘proper ownership by’ such an agent, so as to avoid giving hostages to internalist prejudice. But either way, the point of all these considerations is really to somehow *tie* the candidate process to a given agent).

Given this emerging picture of unified neural underpinnings, and given the rather clear way in which (properly owned or coupled) non-neural resources might nonetheless work closely with the neural regime to further the larger aim of reducing informational surprise, we are back at an impasse. Do we let the unifying neural story carry the load and use that to factor the hybrid cases into a cognitive and non-cognitive component, no matter how ‘well-owned’, intelligent-performance-enhancing, and densely dovetailed they become? Or do we see the further facts about ownership, enhancement, and dovetailing as providing sufficient reason to treat some of the hybrid wholes as realizing new extended cognitive processes in their own right?

It is now over a decade since this question started being debated within philosophy and cognitive science. My current view (arising from the ongoing debates with critics such as Adams and Aizawa, and Rob Rupert, and influenced also by the rich and compelling recent treatment in Sprevak (forthcoming)) is that this debate, though scientifically important, and able to be scientifically informed, looks increasingly unlikely to admit of straightforward scientific resolution. Perhaps this is unsurprising. Science reveals the complex web of structure upon which various forms of apparently intelligent response depend. But which bits of that

web deserve to be labeled the realizers of ‘cognitive processes’ and which do not? This is not a question that, as far as I can currently tell, can be resolved by any simple and non-question-begging empirical means. Nor do any of the more familiar philosophical or metaphysical levers so far applied to the discussion yield sufficient traction. If Wheeler is right, and even the appeal to what you might call the ‘metabolically blindfolded’ folk intuitions probed using the parity principle cannot definitively deliver (due to the folk intuitions lacking sufficient grip upon the notion of vehicles rather than contents) then what remains?

Robert Rupert, in his careful, constructive, ongoing critiques of the arguments for cognitive extension, suggests that an effective lever may yet be provided by appeal to the proven value of focusing our cognitive scientific attentions upon persisting integrated systems and their properties, treating all that lies outside the persisting integrated system as no more than a source of inputs and an arena for outputs. This is an important challenge, but before I attempt a reply, it may be worth noticing something that Rupert misses concerning the argument structure of SSM. For much of the first part of his response, Rupert describes the empirical work and ideas rehearsed in the opening chapters of SSM, with a view to showing that such work can in fact be fully accommodated within the more conservative framework that he favors, dubbed HEMC (the Hypothesis of Embedded Cognition). HEMC states that fully organismically-realized cognitive systems are potently embedded in their local environments. But it was no part of the SSM agenda to claim that *only* the extended perspective could accommodate the empirical results presented in the first few chapters. As mentioned in

the Précis, SSM (as its subtitle “Embodiment, Action, and Cognitive Extension” was meant to indicate) was a book with a dual mission, only half of which was to mount a defense of the extended mind. The first part of the mission was to display, organize, and discuss the burgeoning body of work in embodied, environmentally embedded, cognitive science. To that end, I highlighted a few key principles that seemed to me to be useful and non-obvious - for example, ideas about the self-structuring of information flows, the negotiability of our sense of embodiment and location, the complex role of self-produced language in the unfolding of thought, the role of anarchic, self-stimulating loops, and the later discussions of the possible quantification of ‘embodied advantage’.

A good way to think about all those ideas, it seems to me, is as providing the elements of a (partially) new kind of toolkit for approaching the project of understanding the human mind. This toolkit, we may then notice, is one that is much more friendly to the vision of extended cognitive processes than was that simpler, starker, toolkit that dealt almost exclusively in internal codes, models, and operations, all safely located behind the firewalls of transducer and effector systems. Whereas the latter vision encouraged (though, as Rupert rightly insists, it did not force) us to focus our cognitive scientific attentions on the inner arena alone, the new vision and toolbox immediately alerts us to the larger webs of structure that jointly, and in hugely complex ways, empower and enable human thought, action, and reason. This complex web is not, of course, rendered invisible even by the adoption of a fully classical view of inner codes and operations. But the new tools and perspectives help show that the idea of a genuine yet non-brainbound science of

cognitive systems is not unrealistic. It is striking that these are, increasingly, the tools of choice of those who study mind and cognition.

I agree with Rupert that the use of these new tools and perspectives is consistent with the outright rejection of HEC. It is, of course, consistent also with the outright rejection of even the kind of staunchly organism-centered approach that Rupert himself seems to endorse. Indeed, other critics (such as Adams and Aizawa (2008)) themselves draw the line rather differently to Rupert, insisting that cognitive processing, in humans, is restricted to (aspects of) the brain and central nervous system, while Rupert's view allows appropriate non-neural bodily operations to count, just as long as his key criteria of persistingness and integration are met. This is significant. It shows that for Rupert, it is the threatened transitoriness, and/or merely superficial integratedness, of bio-external operations and storage that most strongly determines their failure to count among the realizers of cognitive processes.

Are these conditions both clear and reasonable? It is unclear (as Theiner (ms) nicely argues) exactly how the proposed criteria would fare *internally* should, for example, the inner neural story turn out to be either strongly modular or (worse still, as thus affecting so-called 'central processing too) massively modular. In each case there is a real danger that we might thus confront a lack of sufficient statistical implication of different neural sub-systems in a wide and ongoing range of tasks, thus forcing us to rule that these neural sub-systems, even when actively involved in some processing, are simply not realizing the agent's mental or cognitive states. Instead,

neurally realized though they might be, they would just count as occasional sources of input to the true – but shrunken, perhaps almost vanishing – engine of cognition. This kind of threat (the threat of unduly shrinking the inner realizers of mind so as to block potential outer realizers) should make us suspicious of the strategy itself.

Why, then, should we suppose that persistingness and this kind of statistically-determined measure of integration are what matters in the first place? Imagine some kind of newly discovered biological creature with a rather complex and environmentally exploitative life-cycle. As this creature makes its way through life, it grows and loses a variety of structural elements. Wide varieties of forms of legs, grippers, wings, eyes, ears, and more, all come and go in wave after wave of epigenetic flux. Nor is what comes and goes any simple function of a developmental program. Rather, this creature is adapted so as to develop such efflorescences partly in response to the whims and fancies of shifting environmental fortune. Bits of tree, metal, sap, plastic, neoprene, and stone, are all fair game as seedcore for the newly emerging bodily forms and structures, which then persist or decay according to need, use and the vagaries of enabling metabolism. Let's name this strangely engaging body-chameleon *Metamorpho*, after that old comic strip character whose ever-shifting surface form it somewhat echoes.

How should we think of *Metamorpho*? First of all, it seems clear (to me, but then I already believe in the extended mind!) that the various come-and-go appendages etc are –despite Rupert's rapid dismissal of the 'growing and shrinking' option in regard to the mental –usefully seen as proper, though

transitory, parts of Metamorpho's bodily form. This is true *despite* the fact that we could, as Rupert might at that point insist, tell a different story about the creature's shape and abilities at each moment: a story that factored things out into whatever morphological elements happen to meet the Rupert test of persistingness and maximal-integration and what do not. Let's assume that were we to do so we'd end up isolating a kind of seldom-seen-in-the-wild core trunk-being. The core trunk being is unable to perceive or locomote, but is always nicely poised to make the most of an open-ended set of environmentally determined opportunities to morph into a being who can.

It seems to me that we have two viable ways of looking at Metamorpho, either of which might be indicated according to our own shifting explanatory purposes. For example, suppose there was a genetic problem affecting the morphing ability in some cases. We might then want to target the trunk-being as itself an entity, and one in need of repair. But for most daily purposes, and for understanding much of the nature of the active, thriving, creature currently wiggling its heterogeneous bundle of sensors and effectors before us, we would surely want to treat each temporary incarnation as the current physical agent. Now, you might say, this is all well and good in the case of Metamorpho. For in this imaginary case we are able visually (and in many other ways) to unproblematically perceive and target the temporary unity that matters. It is this perceptible unity, with attendant opportunities for interaction, that really drives our intuitions. There is, for each Metamorpho slice, a perceptible unity that just jumps out at us despite (let's assume) the lack of full long-term integration

and persistingness of the current ensemble of sensors and effectors.

We may now ask what, in the mental case, might correspond to the visual inspection of current form in the physical case? For just as it seems possible that there be *sufficient* here-and-now physical unity without full and persisting physical integration, so it seems possible that there be *sufficient* here-and-now cognitive unity even without full and persisting cognitive integration. The various devices used in the text, from the parity probe, to the considerations about dovetailing and temporal complexity, as well as the appeal to self-stimulating loops featuring transient items such as pen and paper, were all meant as ways of probing and exploring this hard-to-inspect terrain. For of course, the one thing that we are not allowed to do in this argumentative context is to simply *assume* that the required mental unity is some direct function of biological or organismic unity. Clearly, we are not (not yet at any rate) much like Metamorpho in terms of our gross bodily form. But that mundane physical fact may be blinding us to the surprising extent to which we *are* like Metamorpho when considered as mental beings. We might be like Metamorpho (Metamento?) in that bits of the encountered and designed world become repeatedly and deeply incorporated into our cognitive routines, persisting or decaying according to need, use and the vagaries of our enabling socio-technological cocoon.

What about Rupert's follow-up worry, that we can always re-parse the cognitive cake, so as to do the same science here whether we see ourselves as such shifting hybrid wholes or not? This is surely no more compelling in this case than it was

in the case of Metamorpho. The best way of dealing with this parsing problem, as I argued in SSM, is to accept that there are two perfectly proper cognitive scientific projects here, one of which aims to explain the processes of recruitment, on-the-spot assembly, and longer-term ‘neural dovetailing’ that enable us to *be* these mental Metamorphos, and one of which takes as its object of study minds like ours as they are ‘in the wild’: hybrid bio-socio-technological minds made up of heterogeneous and shifting sets of components. Why care about these unruly hybrids? Because these are the very minds that moment-to-moment negotiate the problem domains most distinctive of human thought and reason. They are *our* minds, in the most compelling and distinctive sense of the term. To Rupert’s suggestion that there is simply no explanatory advantage in seeing ourselves as extended cognitive systems, I would thus reply that to fail to do so may be to fail to see ourselves for the cognitive wholes that we really are.

That double project, and my mention of his trademark notion of mind ‘in the wild’ brings me, finally, to Ed Hutchins skillfully crafted contribution. Hutchins challenges SSM on two principal, and related, counts, each of which amounts to a kind of perceived failure of nerve or vision on my part. The first challenge concerns my take on the processes of ‘recruitment and assembly’. These were the processes by which shifting subsets of neural and bodily resources are brought into fruitful interplay with shifting sets of extra-biological structures, creating new temporary cognitive wholes that are (I claimed) organism-centered without being organism-bound. In describing affairs thus I have, Hutchins suggests, been too concessive, allowing the proponents of

brain-bound cognitive theorizing to skew my own agenda too far in the direction of a more ‘vanilla’ cognitive science. Not only is this seen by Hutchins as an unwanted concession, it is also seen –and this is the second challenge mentioned above– as generating a kind of theoretical vacuum within the treatment as a whole, such that “accounting for the organization of ecological assemblies is the central and unsolved problem of the book”.

Hutchins goes on to make and pursue his own powerful positive suggestion, which is that much of the apparently ‘missing’ work is actually done by our own slowly evolved and variously transmitted cultural practices. The idea is thus that it is these practices (see also Hutchins (2008)) that bear much of the explanatory weight, as far as those processes of on-the-spot recruitment and assembly are concerned. As such it is both unnecessary and incorrect to depict the processes of recruitment and assembly as being heavily brain-based (and the complex self-stimulating cycles they generate thus ‘organism-centered’) in supposed contrast to the resulting - more distributed and transient - wholes. The ‘enculturated supersized mind’, if Hutchins is correct, simply does not *need* to solve those ‘hard problems’ of on-the-spot recruitment and ecological assembly. For the most part, our predecessors did that for us, and courtesy of the cumulative culturally-encoded fruits of their labors we (the theorists) can now make the final mental flip needed to complete the revolution, arriving at a view of mind stripped at last of every vestige of the ‘scaffolding of brainbound thinking’.

This is an inspiring, important, and ambitious vision, and I have much to learn from it. But at the same time, I do think it

is important not to *undervalue* the role, in the generation and maintenance of the many hybrid forms of human mentality, of the (currently) unique and critical core contribution made by that remarkable organ, the human brain. Brains *are* special, and to assert this need mark no slippery-slope concession to good old-fashioned internalism as an account of mind. It is fully consistent with thinking (as I do) that Hutchins is absolutely right to stress the major role of transmitted cultural practices in *setting the scene* for various neurally-based processes of cognitive assembly. To see what I mean by this, we need only remind ourselves that successful cognitive assembly is itself a product of many processes operating over very different timescales. I would not want to deny, for example, that the cultural practices of pen-and-paper based long multiplication set the scene by providing me with both a pre-structured recipe for success, a well-honed cultural practice (schooling) to help me benefit from that recipe, and a pre-selected set of supporting materials and structures (pen, paper) all ripe for assembly into a new problem-solving whole. The contributions of the cultural backdrop are thus truly profound. But they should not blind us to the amazing potency of the human brain in enabling me, in various ways and at various times, to profit from that prodigious cultural provision.

It is not, of course, any part of Hutchins aim to downplay the role of the brain in human affairs! But in asserting, in effect, that the appeal to cultural practices is *sufficient* to account for all the crucial work of cognitive assembly, I think Hutchins is failing to attend to important differences concerning the shape and timescale of the processes concerned. My own targets, in the discussions in SSM of cognitive assembly, were

the processes operating in the here-and-now. They were the processes whose overall effect is to tie together a set of information-processing resources (some might be neural, some bodily, some bio-external) in delicate temporal harmonies, orchestrating calls to external information stores, calls to internal information stores, neural transformations, and a variety of externally-mediated transformations, in the ways necessary for that whole hybrid ensemble to get to grips with some problem. Now it may well be true that some of this load is sometimes borne by the skilled performances of others and hence that there can be, as Hutchins nicely shows, a very interesting social dimension even to on-the-spot recruitment and assembly. But even here, it is still individual biological brains (though working together in these cases) that are, in the here-and-now, the most active orchestrating elements in this process.

It is crucial to the story I am telling that the biological brain adapts, selects, and alters, its own internal routines so as more and more fluently to exploit the reliable presence of all those specific, culturally selected, tuned, and delivered, resources. For it is only in that way that we achieve the kind of complex temporally nuanced dovetailing (between the neural and the rest) that warrants treating a temporary ensemble, Metamorpho-like, as a new, genuine, cognitive whole. There is no conflict, as far as I can see, between my claim that the biological brain is the essential core element that allows all this dovetailing and assembly to take place, and Hutchins' claim that much of the explanatory burden (in any given case of ecological assembly) is borne by long, hard-won, chains of cultural innovation and transmission. Both claims are true

and important, but they target differing timescales and processes of adaptation and change.

Hutchins' primary cultural objects are, of course, shared human practices rather than simply collections of artifacts or materially transmitted recipes. Perhaps making this final perspective flip, into collaborative human action structured by the existing practices of the various human groups in which we participate, goes further towards absolving the individual human brain of the bulk of (what I see as) the burden of ecological assembly. Unfortunately (for me) I don't yet see, in any detail, quite how this can be so. For as Hutchins himself says, it is only the 'special super-flexible medium' of the brain that *allows* such shared practices to come to orchestrate human learning and response in the first place. In depicting the processes of on-the-spot recruitment and exploitation as neurally-centered, I meant only to stress the pivotal role, on all these shorter time-scales, of the specifically neural changes that immersion in those cultural practices presumably inculcate.

Hutchins response might be that we should simply reject the conceptual separation between the processes operating on these various timescales. That is how I read his key suggestion that "both the constraints of cultural practices and the malleable internal microdemons can be seen as elements of a single adaptive system". But while I agree that these are indeed (also) elements of a single long-term adaptive system, that does nothing to diminish the conceptual separation between the long-term evolution of cultural practices, the medium-term effects of my immersion in such practices, and the short-term processes by means of which my brain then

participates in what (from an extended mind perspective at least) are new hybrid cognitive routines that productively criss-cross brain, body, and world.

I must, however, plead guilty as charged to prolonged and continuing neglect of the massive social and cultural dimensions that shape and enable our actual cognitive practices. Perhaps that failure of attention leads me to overvalue, even on the smaller time-scales, the ‘maintaining-and-orchestrating’ contributions of the biological brain. Thus just as some linguists (e.g. Christiansen and Kirby (2003) Kirby et al (2008)) now believe that public languages have, via iterated processes of trans-generational learning and transmission, progressively fitted themselves to brains like ours (so that it is mostly the languages that have ‘learnt’ about brains like ours rather than the other way around) so it may be that other forms of cultural practice and device have done likewise, so as slowly to become the kinds of object or practice towards which massive neural dovetailing is simply not required. Such objects would simply be extraordinarily ‘fit to be assimilated’ by brains like Determining the precise nature, extent, and distribution of here-and-now neural labor in the orchestration and maintenance of hybrid processing ensembles is thus a complex empirical matter, and one that almost certainly has no unique solution.

Where does all this leave the saga of SSM and the extended mind? I am now fairly convinced (for a good argument here, see Sprevak (in press)) that there will be no straightforward empirical resolution to the questions concerning cognitive extension. I am convinced, too, that the perspective that views some cognitive processes as looping through brain,

body, and world will continue to be productive, both as a source of philosophical stimulation and scientific insight. This is because we *still* pay too little attention, in both science and philosophy (though things are changing on both fronts) to the massive role that self-stimulating loops and the active structuring of information flows play in the construction and unfolding of human thought and reason. What the near-future will bring, I strongly suspect, is a much better grip on the distinctive contribution of the brain within such potent loops and cycles. By locating that story in a wider framework that displays the neural contribution as itself a manifestation of a larger imperative (to structure brain, body, world, and action in ways that work together to reduce informational surprise) we may yet reveal the true nature of that murky family business in which brain, body, world, and action so potently conspire.

References

- Adams, F. and Aizawa, K. (2001). The Bounds of Cognition. *Philosophical Psychology* 14:1: 43-64.
- Adams, F. and Aizawa, K. (2008). *The Bounds of Cognition* (Blackwell, MA)
- Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997) 'Deictic codes for the embodiment of cognition'. *Behavioral and Brain Sciences*, 20, 723-767
- Christiansen, M. H. and Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, 7(7) :300-307.
- Clark, A and Chalmers, C (1998) The Extended Mind *Analysis* 58:1:7-19
- Dennett, D (1991) *Consciousness Explained* (Little Brown, Boston)
- Dennett, D (1996) *Kinds of Minds* (Basic Books, NY)

Donald, M (1991) *Origins of the Modern Mind* (Harvard University Press, Cambridge, MA)

Friston, K (2005). "A theory of cortical responses." *Philosophical Transactions: Biological Sciences* 369(1456): p. 815 -836.

Friston K. (2009) The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences* 13:7: p. 293-301

Friston K (2010) The free-energy principle: a unified brain theory? *Nature Reviews: Neuroscience*. 11:2: p. 127-38

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review* 113(3) 461-482.

Haugeland, J (1998) "Mind Embodied and Embedded" in J. Haugeland *Having Thought: Essays in the Metaphysics of Mind* (Harvard University Press, Cambridge, MA) 207-240.

Hohwy, J. (2007) Functional integration and the mind. *Synthese* 159(3): 315-328

Hurley, S. (1998) *Consciousness in Action*. (Cambridge, MA: Harvard)

Hutchins, E (1995) *Cognition In The Wild* (MIT Press: Camb. MA)

Hutchins, E (2008) The role of cultural practices in the emergence of modern human intelligence. *Phil. Trans. R. Soc. B* 363, 2011-2019.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative Cultural Evolution in the Laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681-10686.

Kirsh, D and Maglio, P (1994) On Distinguishing Epistemic From Pragmatic Action *Cognitive Science* 18:513-549

Lee, T.S., Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America, A*. 20(7): 1434-1448

Lungarella, M. and Sporns, O. (2005). Information self-structuring: key principles for learning and development. *Proceedings 2005 IEEE Intern. Conf. Development and Learning*, pp. 25-30.

Lungarella, M., Pegors, T., Bulwinkle, D, and Sporns, O. (2005). Methods for quantifying the information structure of sensory and motor data. *Neuroinformatics*, 3(3):243-262.

Menary, R (2007) *Cognitive Integration: Attacking The Bounds of Cognition*. Palgrave Macmillan

Milner, A and Goodale, M (1995) *The Visual Brain in Action* (Oxford University Press, Oxford, UK)

Milner, D. and Goodale, M. (2006)'Epilogue: Twelve Years On" in Milner, D and Goodale, M *The Visual Brain in Action: second edition* Oxford: Oxford University Press p.207-252

Noë, A. (2004) *Action in Perception*. Cambridge, MA: The MIT Press.

Rowlands, M (1999) *The Body in Mind: Understanding Cognitive Processes* (Cambridge University Press, Cambridge, UK)

Rowlands, M (2006) *Body Language: Representing in Action*, MIT Press

Sprevak, M (forthcoming) Inference to the hypothesis of extended cognition. *Studies in History and Philosophy of Science*

Sutton, J (In Press) 'Batting, Habit, and Memory: the embodied mind and the nature of skill' *Sport in Society*

Theiner, G (ms) The HEC Strikes Back: Extended Cognition and Rupert's 'Priority of Cognitive Systems'.

Wheeler, M (2005) *Reconstructing the Cognitive World* (MIT Press, Camb. MA)

Wilson, R. A. (1994) "Wide Computationalism" *Mind* 103:351-372

Wilson, R. A. (2004) *Boundaries of the Mind: The Individual in the Fragile Sciences--Cognition* (Cambridge University Press, Cambridge, UK)

¹ This idea is introduced in Clark and Chalmers (1998)

² Clark and Chalmers (1998). See also, among others, Dennett (1991) (1996), Donald (1991), Haugeland (1998), Hurley (1998), Hutchins (1995), Menary (2007), Noë (2004), (2009), Rowlands (1999) (2003) (2006), Sutton (In Press), Wheeler (2005), Wilson (2004).