"Memento's Revenge: Objections and Replies to the Extended Mind" to appear in R. Menary (ed) PAPERS ON THE EXTENDED MIND (press?)

Memento's Revenge: The Extended Mind, Extended.

Andy Clark

> In the movie, *Memento,* the hero, Leonard, suffers from a form of anterograde amnesia that results in an inability to lay down new memories. Nonetheless, he sets out on a quest to find his wife's killer, aided by the use of notes, annotated polaroids, and (for the most important pieces of information obtained) body tattoos. Using these resources he attempts to build up a stock of new beliefs and to thus piece together the puzzle of his wife's death. At one point in the movie, a character exasperated by Leonard's lack of biological recall, shouts:
>
> "YOU know? What do YOU know. YOU don't know anything. In 10 minutes time YOU won't even know you had this conversation"

Leonard, however, believes that he does, day by day, come to know new things. But only courtesy of those photos, tattoos, tricks and ploys. Who is right?

These are the kinds of question addressed at length in the paper (co-authored with David Chalmers) 'The Extended Mind'. Is the mind contained (always? sometimes? never?) in the head? Or does the notion of thought allow mental processes (including believings) to inhere in extended systems of body, brain and aspects of the local environment? The answer, we claimed, was that mental states, including states of believing, could be grounded in physical traces that remained firmly outside the head. As long as a few simple conditions were met (more on which below), Leonard's notes and tattoos could indeed count as new additions to his store of long-term knowledge and dispositional belief.

In the present treatment I revisit this argument, defending our strong conclusion against a variety of subsequent observations and objections. In particular, I look at objections that rely on a contrast between the (putatively) intrinsic content of neural symbols and the merely derived content of external inscriptions, at objections concerning the demarcation of scientific domains via natural kinds, and at objections concerning the ultimate locus of agentive control and the nature of perception versus introspection. I also mention a possible alternative interpretation of the argument as (in effect) a reductio of the very idea of the mind as an object of scientific study. This is an interesting proposal, but one whose full evaluation must be left for another time.

First, though, it will help to briefly review the original argument from Clark and Chalmers (1998).

1. Tetris and Otto.

Two examples animated the original paper. The first involved a human agent playing the arcade game TETRIS. The human player has the option of identifying the falling pieces (a) by mental rotation or (b) by the use of the onscreen button that causes the falling zoid to rotate. Now imagine (c) a future human with both normal imaginative rotation capacities and also a retinal display that can fast-rotate the image on demand, just like using the rotate button. Imagine too that to initiate this latter action the future human issues a thought command straight from motor cortex (ie this is the same technology as actually used in so-called thought control experiments- see eg Graham-Rowe (1998)).

Now let us pump our intuitions. Case (a) looks, we argues, to be a simple case of mental rotation. Case (b) looks like a simple case of non-mental (merely external) rotation. Yet case (c) now looks hard to classify. By hypothesis, the computational operations involved are the same as in case (b). Yet our intuitions seem far less clear. But now add the Martian player (case 4) whose natural cognitive equipment includes (for obscure ecological reasons) the kind of bio-technological fast-rotate machinery imagined in case (3). In the Martian case, we would have no hesitation in classifying the fast-rotations as species of mental rotation.

With this thought experiment as a springboard, we offered a Parity Principle as a rule of thumb viz:

Parity Principle.

> 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.
> (from Clark and Chalmers (1998) p.XX)

The Parity Principle invites us to treat the players' use of the external rotate button, the cyberpunk implant, and the Martian native endowment as all on a cognitive par. But of course there are differences. Most strikingly, in case (2) the fast-rotate circuitry is located outside the head and the results are read-in by perception, whereas in cases (3) and (4) the circuitry is all bounded by skin and skull and the results are read-off by introspection. I return to these issues below. Nonetheless there remained, we argued, at least a prima facie case for parity of treatment based on the deep computational commonalities rather than simple prejudices about skin and skull, inner and outer. The most important difference, we felt, concerned not the arbitrary barriers of skin and skull, or the delicate (and potentially question-begging) call between perception and introspection, but the more basic functional issues of portability and general availability for use. The standard player's use of the fast-rotate button is limited by the availability of the Tetris console,

whereas the cyberpunk and Martian players exploit a resource that is part of the general equipment with which they confront the world.

Taking the argument one step further, we then considered a second example, one designed to address the portability issue and to extend the treatment to the more central case of an agent's beliefs about the world. This was the case of Otto and Inga.

Inga hears of an intriguing exhibition at MOMA (the Museum of Modern Art in New York). She thinks, recalls it's on 53rd St, and sets off. Otto suffers from a mild form of Alzheimer's, and as a result he always carries a thick notebook. When Otto learns useful new information, he always writes it in the notebook. He hears of the exhibition at MOMA, retrieves the address from his trusty notebook and sets off. Just like Inga, we claimed, Otto walked to 53rd St. because he *wanted* to go to the museum and *believed* (*even before consulting his notebook*) that it was on 53rd St. The functional poise of the stored information was, in each case, sufficiently similar (we argued) to warrant similarity of treatment. Otto's long-term beliefs just weren't all in his head.

In the paper we showed, in detail, why this was not equivalent to the more familiar Putnam/Burge style externalism, arguing that what was at issue was more like an environmentally extended case of narrow content than a case of broad content. The idea was that the causally active physical vehicles of content and of cognitive processes could be spread across the biological organism and the world. This

was quite different, we claimed, from any form of passive, reference-based externalism.

Further, we allowed that (as far as our argument was concerned) conscious mental states might well turn out to supervene only on local processes inside the head. But insofar as the scope of the mental is held to outrun that of conscious, occurrent contents (to include, for example, my long-term dispositional beliefs as well as my current conscious believings) there was no reason to restrict the physical vehicles of such non-conscious mental states to states of the brain or central nervous system.

In response to the more serious (in our opinion) concerns about availability and portability, we offered a rough-and-ready set of additional criteria to be met by non-biological candidates for inclusion into an individual's cognitive system. They were:

> 1. That the resource be reliably available and typically invoked.
> (Otto always carries the notebook and won't answer that he 'doesn't know' until after he has consulted it).
>
> 2. That any information thus retrieved be more-or-less automatically endorsed. It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed about as trustworthy as something retrieved clearly from biological memory.

3. That information contained in the resource should be easily accessible as and when required.

Applying the three criteria yielded, we claimed, a modestly intuitive set of results for putative individual cognitive extensions. A book in my home library would not count. The cyberpunk implant would. Mobile access to Google would not (it would fail condition (2). Otto's notebook would. Other people typically would not (but could in rare cases), etc.

There is one reply which we consider in the paper that I choose to repeat here, just because it is still the most common response to our story. I call it the Otto 2-step and it goes like this:

> "all Otto actually believes (in advance) is that the address is in the notebook. That's the belief (step 1) that leads to the looking (step 2) that then leads to the (new) belief about the actual street address"

Despite its initial plausibility, we do not think this can work. Suppose we now ask why we do not depict Inga in similar terms? Why don't we say that Inga's only antecedent belief was that the information was stored in her memory, and depict her retrieval as an Inga 2-step?

Intuitively, the reason seems to be that in the case of Inga, the 2-step model adds spurious complexity: "Inga wanted to go to MOMA. She

believed that her memory held the address. Her memory yielded 53rd St. ...". What's more, it seems likely that in the normal course of events Inga relies on no beliefs about her memory as such. She just uses it, transparently as it were. But *ditto* (we may suppose) for Otto: Otto is so used to using the book that he accesses it automatically when bio-memory fails. It is transparent equipment for him just as biological memory is for Inga. And in each case, it adds needless and psychologically unreal complexity to introduce additional beliefs about the book or biological memory into the explanatory equations.

In the paper we consider a few variants on this theme, but all go the same way in the end. Inga's biological memory systems, working together, govern *her* behaviors in the functional ways distinctive of believing. Otto's bio-technological matrix (the organism and the notebook) governs his behavior in the same sort of way. So the explanatory apparatus of mental state ascription gets an equal grip in each case and what looks at first like Otto's action (looking up the notebook) emerges as part of Otto's thought. Mind, we conclude, is congenitally predisposed to seep out into the world.

2. Intrinsic Content

Adams and Aizawa (2001) present a variety of considerations meant to undermine a position that they dub 'transcranialism' viz the view that " cognitive processes extend in the physical world beyond the bounds of the brain and the body"(op cit 43). This is a view that they associate, in varying degrees, with the work of Merlin Donald, Daniel

Dennett, Ed Hutchins and Clark and Chalmers. While conceding that transcranialism is "logically and nomologically possible" (and might thus be true of, for example, some alien species on a different planet) it is, they maintain, false in the case of human cognition. They thus opt for a "contingent intracranialism about the cognitive" (op cit 43).

Top of their list of reasons for this oddly mixed judgement is that in the human case (though not, presumably, in some imaginable alien case) the external media (Adams and Aizawa focus almost entirely on simple external symbolic media such as Otto's notepad) support only *derived* content. Inner symbols, on the other hand, are said to have *intrinsic* content. Thus we read that:

> "strings of symbols on the printed page mean what they do in virtue of conventional associations….The representational capacity of orthography is in this way derived from the representational capacities of cognitive agents. By contrast the, cognitive states in normal cognitive agents do not derive their meanings from conventions or social practices…"(48)
>
> And later on that:
>
> "Whatever is responsible for non-derived representations seems to find a place only in brains" (63)

Suppose we grant, for the sake of argument, something I am actually fundamentally inclined to doubt, viz, that there is a clear and distinct

sense in which neural representations get to enjoy 'intrinsic contents' of some special kind, quite unlike the kinds of content that figure in external inscriptions. The most obvious way to unpack this is, still following Adams and Aizawa, in terms of a fundamental distinction between inscriptions whose meaning is conventionally determined and states of affairs (eg neural states) whose meaning-bearing features are not thus parasitic. The question is, must everything that is to count as part of an individual's mental processing be composed solely and exclusively of states of affairs of this latter (intrinsically content-bearing) kind? I see no reason to think that they must.

For example, suppose we are busy (as part of some problem-solving routine) imagining a set of Venn Diagrams/ Euler Circles in our mind's eye? Surely the set-theoretic meaning of the overlaps between say, two intersecting Euler circles is a matter of convention? Yet this image can clearly feature as part of a genuinely cognitive process.

To this, Adams and Aizawa might reply as follows: "Ah but the image, when understood, must be triggering neural goings-on with intrinsic content: and it is in that that the understanding eventually consists" But so what? When Otto reads the notebook, neural goings-on with intrinsic content are likewise triggered. To which (perhaps) the reply: "OK, but what about before that, when the inscription is simply in the notebook? Surely Inga's stored beliefs must continuously have intrinsic content too, not just her occurrent ones"

Now this is a harder question, and one which might even begin to suggest the ultimate fragility of the very idea of intrinsic content. But we can sidestep that discussion with a simple thought experiment that builds on the original Parity Principle rehearsed in section 1. What if we found Martians whose biological routines stored *bit-map images* of printed words that they could later access (and interpret) via bit-mapped signals sent to visual cortex? Surely we would have no hesitation in embracing that kind of bit-mapped storage as part of the Martian system? It is not unlike, in fact, the case of those human memory masters who are able to recall a passage from a text by first recalling, then imaginatively inspecting, a photo-like image of the original page.

In the light of all this, the fair demand is (at most) that we should somehow link those stored representations whose contents are derived (conventional) to ones whose contents, at least when occurrent, are 'intrinsic' (by whatever standards of intrinsic-ness Adams and Aizawa imagine may prevail). But such linking can be (and is) routinely achieved for representations stored outside the head. The inscriptions in Otto's notebook, I conclude, can be properly poised in any larger cognitive economy that includes states with intrinsic content.

In fact, after a long discussion of all this, Adams and Aizawa actually concede that:

> "Having argued that, in general, there must be non-derived content in cognitive processes, it must be admitted that it is unclear to what extent every cognitive state of each cognitive process must involve non-derived content" (50).

At which point there is really no case (concerning intrinsic content) left to answer.

3. Scientific Kinds and Functional Similarity.

In the same paper, Adams and Aizawa also raise a very different kind of worry. This concerns the nature and feasibility of the scientific enterprise implied by taking transcranialism seriously. The worry, in its simplest form, is that "science tries to carve nature at its joints" (51). But (they argue) the various types of neural and extra-neural goings-on that the trancranialist lumps together as 'cognitive' seem to have little or nothing in common by way of underlying causal processes. The causal arrangements whereby external stuff contributes to considered action look to be very different to those whereby internal stuff does. As a result, the argument continues, there can be no unified science of the extended mind. Better, then, to keep the domains apart and settle for a unified science of the inner

(properly mental) goings-on, and another science (or sciences) of the (non-mental) rest.

To make this concrete, we are invited to consider the process that physically rotates the image on the Tetris screen. This, they correctly note, is nothing like any neural process. It involves firing electrons at a cathode ray tube! It requires muscular activity to operate the button. Similarly, "Otto's extended 'memory recall' involves cognitive-motor processing not found in Inga's memory recall."(55) And so on. More generally, they suggest, just look at the range of human memory augmenting technologies (photo albums, tattoos (for Memento), rolodexes, palm pilots, notepads etc:

> "what are the chances of their being interesting regularities that cover humans interacting with all these sorts of things? Slim to none, we speculate" (61)

By contrast, biological memory systems are said to:

> "display a number of what appear to be law-like regularities, including primacy effects, recency effects, chunking effects and others" (61).

And unlike the biological memory processes:

> "transcranial [extended] processes are not likely to give rise to interesting scientific regularities. There are no laws covering

> humans and their tool-use over and above the laws of intercranial [inner] human cognition and the laws of the physical tools"(61)

The first thing to say in response to all this is that it is unwise to judge, from the armchair, the chances of finding 'interesting scientific regularities' in any domain, be it ever so superficially diverse. Consider, for example, the recent successes of complexity theory in unearthing unifying principles that apply across massive differences of scale, physical type, and temporality. There are power laws, it now seems, that compactly explain aspects of the emergent behavior of systems ranging from XX to YY. In a similar vein, it is quite possible that despite the bottom-level physical diversity of the processes that write to, and read from, Otto's notebook, and those that write to, and read from, Otto's biological memory, there is a levcel of description of these systems that treats them in a single unified framework (for example, how about a framework of information storage, transformation and retrieval!). The mere fact that Adams and Aizawa can find ONE kind of systemic description at which the underlying processes look wildly different says very little, really, about the eventual prospects for an integrated scientific treatment. It is rather as if an opponent of rule and symbol models of mental processing were simply to cite the deep physical differences between brains and Von Neumann computers as proof that there could be no proper science that treated processes occurring in each medium in a unified way. Or, to take a different kind of case, as if one were to conclude

from the fact that chemistry and geology employ distinct vocabularies and techniques, that the burgeoning study of geochemistry is doomed from the outset. But neither of these, I presume, are conclusions that Adams and Aizawa would wish to endorse.

The bedrock problem thus lies with the bald assertion that "the cognitive must be discriminated on the basis of underlying causal processes" (op. cit 52). For it is part of the *job* of a special science to establish a framework in which superficially different phenomena can be brought under a unifying explanatory umbrella. To simply cite radical differences in some base-level physical story goes no way at all towards showing that this cannot be done. Moreover, it is by no means clear that acceptable forms of unification require that all the systemic elements behave according to the same laws. As long as there is an intelligible domain of convergence, there may be many sub-regularities of many different kinds involved. Think, for example, of the multiple kinds of factor and force studied by those interested in creating better home audio systems. Even if 'home audio' is rejected as any kind of unified science, it certainly names a coherent and proper topic of investigation. The study of mind might, likewise, need to embrace a variety of different explanatory paradigms whose point of convergence lies in the production of intelligent behavior.

It is quite possible, after all, that the *inner* goings-on that Adams and Aizawa take to be paradigmatically cognitive themselves will turn

out to be a motley crew, as far as detailed causal mechanisms go, with not even a family resemblance (at the level of actual mechanism) to hold them together. It is arguable, for example, that conscious seeing and non-conscious uses of visual input to guide fine-grained action, involve radically different kinds of computational operation and representational form. (REF Milner and Goodale). And (Adams and Aizawa to the contrary) some kinds of mental rehearsal (such as watching sports, or imagining typing a sentence) do seem to re-invoke distinct motor elements, while others (imagining a lake) do not. (Decety and Grezes (1999)). Some aspects of biological visual routines even use a form of table look-up (PS Churchland and T Sejnowski (1992).

In the light of all this, my own suspicion is that the differences between external-looping (putatively cognitive) processes and purely inner ones will be *no greater than those between the inner ones themselves.* But insofar as they all form parts of a flexible and information-sensitive control system for a being capable of reasoning, of feeling, and of experiencing the world (a 'sentient informavore' if you will) the motley crew of mechanisms have something important in common. It may be far less than we would require of any natural or scientific kind. But so what?

The argument-from-scientific-kinds is thus doubly flawed. It is flawed in virtue of its rather limited conception of what makes for a proper scientific or explanatory enterprise. And it is flawed in its

assessment of the potential for some form of higher-level unification despite mechanistic dissimilarities. It is, above all else, a matter of empirical discovery, not armchair speculation, whether there can be a fully-fledged science of the extended mind.

It is also perhaps worth noting that nascent forms of just such a science have been around for quite some time. The field of HCI (human-computer interaction) and its more recent cousins HCC (human-centered Computing) and HCT (human-centered Technologies) are ongoing attempts to discover unified scientific frameworks in which to treat processes occurring in (and between) biological and non-biological information-processing media (see, for example, Norman (1999) ,Rogers et al (2003)). Likewise, the existence of academic bodies such as the Cognitive Technology Society (and their excellent new journal) likewise attests to the viability of the attempt (though it is, of course, no guarantee of ultimate success) to understand minds and technologies as aspects of an integrated whole.

Adams and Aizawa try to parlay the misconceived appeal to scientific kinds into a kind of dilemma. Either (the argument goes) Clark and Chalmers are radically mistaken about the causal facts or (more likely) they are closet behaviorists. On the one horn, if our claim is that "the active causal processes that extend into the environment are just like the ones found in intracranial cognition" (56) we are just plain wrong. On the other horn, if we don't care about that, and claim only that "Inga and Otto use distinct sets of

capacities in order to produce similar behavior" (56) then we are behaviorists.

This is surely a false dilemma. To repeat, our claim is not that the processes in Otto and Inga are identical, or even similar, in terms of their detailed implementation. It is simply that, in respect of the role that the long-term encodings play in guiding current response, both modes of storage can be seen as supporting dispositional beliefs. It is the way the information is poised to guide reasoning (such as conscious inferences that nonetheless result in no overt actions) and behavior that counts. This is not behaviorism but functionalism. It is systemic role that matters, not brute similarities in public behavior (though the two are of course related). Perhaps Adams and Aizawa believe that functionalism just *is* a species of behaviorism. If so, we plead guilty to the charge but find it less than damning.

A related concern has been raised (personal communication) by Terry Dartnall. Dartnall worries that the plausibility of the Otto scenario depends on an outmoded image of biological memory itself: the image of biological memory as a kind of static store of information awaiting retrieval and use. This image, Dartnall claims, cannot do justice to the active nature of real memory. It is somewhat ironic, Dartnall adds, that the present author (in particular) should succumb to this temptation, given his long history of interest in, and support for, the connectionist alternative to classical (text and rule based) models of neural processing. By way of illustration (though the illustration may actually raise other issues too, as we shall see) he

offers the following example: suppose I have a chip in my head that gives me access to a treatise on nuclear physics. That doesn't make it true that *I know* about nuclear physics. In fact, the text might even be in a language I don't understand. 'Sterile text', Dartnall concludes, cannot support cognition (properly understood). In a sense, then, the claim is (once again) that text-based storage is so unlike biological memory that any claim of role-parity must fail.

This is an interesting line of objection but one that ultimately fails for reasons closely related to the discussion of intrinsic content in section 1. Certainly, biological memory is an active process. And retrieval is to a large extent reconstructive rather than literal: what we recall is influenced by our current mood, our current goals, and by information stored after the time of the original experience (REFS eg Roediger). It is possible, in fact, that biological memory is such an active process as to blur the line between memory systems and reasoning systems. All this I happily accept. But to repeat, our claim is not (ridiculously) that the notebook considered alone would constitute any kind of cognitive system. It would not, but in this respect it is no worse off than a single neuron, or neural population. Rather, the claim is that in the special context of the rest of Otto's information-processing economy, the notebook is co-opted into playing a real cognitive role. And the informal test for this is, just supposing some inner system provided the functionality that Otto derives from the reliable presence of the notebook, would we hesitate to classify that inner system as part of Otto's cognitive apparatus?

The reader must here rely on her own intuitions. But ours are clear. There would be no such hesitation. To cement the intuition, I considered (section 1) the Martian's with their additional bit-map memories, or humans with quasi-photographic recall. To add one case to the pot, consider now the act of rote-learning. When we learn a long text by rote, we create a memory object that is in many ways unlike the standard case. For example, to recall the sixth line of the text we may have to first rehearse the others. Moreover, we can rote-learn a text we do not even understand (eg a Latin text, in my case). Assuming that we count rote learning as the acquisition of some kind of knowledge (even in the case of the Latin text) it seems that we should not be bothered by the consequences that Dartnall unearths. The genuine differences that exist between the notebook-based storage and standard cases of biological memory do not matter, since our claim was not one of identity in the first place.

The question is, how to balance the Parity Principle (which makes no claims about process-level identity at all, and merely identifies a state or process as cognitive) against the somewhat stronger claim of 'sufficient functional similarity' that underpins treating Otto's notebook as a contributor to Otto's long-term store of dispositional beliefs? But the answer emerges as soon as we focus on the role the retrieved information will play in guiding current behavior. It is at that point (and there, of course, all kinds of active and occurrent processing come into play as well) that the functional similarity becomes apparent. True, that which is stored in Otto's notebook won't shift and alter while stored away. It won't participate in the

ongoing underground reorganizations, interpolations, and creative mergers that characterize much of biological memory. But *when called upon*, its immediate contributions to Otto's behavior still fit the profile of a stored belief. Information retrieved from the notebook will guide Otto's reasoning and behavior in the same way as information retrieved from biological memory. The fact that WHAT is retrieved may be different is unimportant here. Thus had Otto stored the information about the color of the car in the auto accident in biological memory, he may be manipulated into a false memory situation by a clever experimenter. The notebook storage is sufficiently different to be immune to that manipulation (though others will be possible). But the information recalled (veridical in one case but not the other) will nonetheless guide Otto's behavior (the way he answers questions and the further beliefs he forms etc) in exactly the same kind of way.

As a final thought hereabouts, reflect that for many years the classical 'text and rule based' image of human cognition was widely accepted. During that time, no-one (to my knowledge) thought that an implication of this was that humans were not cognizers! It might have turned out that all our memory systems operated as sterile storage, and that false memory cases etc were all artifacts of retrieval processes. This shows, again, that there is nothing intrinsically 'non-cognitive' about less active forms of storage.

There is, however, a much bigger issue bubbling beneath the surface of this last discussion. It is the question of how to extend the notion of cognition and cognitive processes beyond the normal human case. Should we fix the domain of the cognitive by reference to the actual (detailed) processing profiles of normal human agents (deferring, I suppose, to our best final science of the normal human brain)? Or should we count ourselves as already commanding an understanding capable of extension to new cases? The argument by Clark and Chalmers assumes that we do possess some such understanding, and that it is rooted, roughly speaking, in our implicit knowledge of the distinctive functional role of cognitive processes in guiding intelligent behavior. It is this knowledge that allows us to count alien processes in non-human animals as properly cognitive, and upon which we must rely when applying the informal test embodied in the Parity Principle. The alternative (making everything depend on identity with processing in the normal human case) strikes us as both anthropocentric and ultimately unworkable. But this is a very large topic indeed and one that I cannot fruitfully pursue much further in the present essay (see the end of section 6 for a few additional comments).

4. On Control

Keith Butler raises the following worry:

> " ... there can be no question that the locus of computational and cognitive control resides inside the head of the subject [and involves] internal processes in a way quite distinct from the way external processes are involved. If this feature is indeed the mark of a truly cognitive system, then it is a mark by means of which the external processes Clark and Chalmers point to can be excluded"
> (Butler (1998), p. 205)

Butler's suggestion is that even if external elements sometimes participate in processes of control and choice (the knot in the hanky , the entry in the notebook) still it is always the biological brain that has the final say, and that here we locate the difference that (cognitively speaking) really makes a difference. The brain is the controller and chooser of actions in a way all that external stuff is not, and so the external stuff should not count as part of the REAL cognitive system.

In fact, there are at least two issues hereabouts. One concerns the functional poise of the neural computations, and the claim that they (alone) are the "locus of computational and cognitive control". The other concerns the nature of the processes, which are said (echoing Adams and Aizawa and Dartnall) to act "in a way quite distinct from the way external processes are involved". I think this latter worry has already been laid to rest. What of the former: the worry about ultimate choice and control?

The worry is interesting because it again highlights the deceptive ease with which critics treat the inner realm itself as scientifically unified. Thus suppose we re-apply the "locus of control" criterion *inside the head*. Do we now count as *not part of my mind or myself* any neural subsystems that are not the ultimate arbiters of action and choice? Suppose only my frontal lobes have the final say - does that shrink the real mind to just the frontal lobes!? What if (as Dan Dennett sometimes suggests, most recently in his (2003)) no subsystem has the 'final say'. Has the mind and self just disappeared?

There is a sense, I think, in which much opposition to the idea of non-biological cognitive extension trades on a deeply mistaken view of the thinking agent as some distinct inner locus of final choice and control. This is a view that I argue against at length in Clark (2003). But for now, let us simply notice that even if there WAS some distinct inner locus of final choosing, there is no reason at all to identify that with the mind or the 'cognitive agent'. Thus my long-term stored knowledge is often called upon in my decision-routines, but the longterm storage itself is no more an ultimate deciding-routine than is Otto's notebook. But (and this is the crunch) to discount all that long-term stored knowledge as partially constitutive of my mind and self is to divorce my identity as an agent from the whole body of memories and dispositional beliefs that guide and shape my behaviors. And this, I maintain, is to shrink the mind and self beyond

recognition, reducing me to a mere bundle of control processes targeted on occurrent mental states.
The argument from ultimate control does not reveal the mark of the mental, or the source of the self.

5. Perception and Development.

A common worry is that the role of perception, in 'reading in' the information from the notebook, marks a sufficient disanalogy to discount the notebook as part of Otto's cognitive apparatus. We made a few brief comments on this issue in the original paper, noting that whether the 'reading-in' counts as perceptual or introspective depends, to a large extent, on how one classifies the overall case. From our perspective the systemic act is more like an act of introspection than one of perception. As a result each side is here in danger of begging the question against the other.

Thus Butler complains that:

> "In the world-involving cases, the subjects have to act in a way that demands of them that they perceive their environment [whereas Inga just introspects] ... the very

> fact that the results are achieved in such remarkably different ways suggests that the explanation for one should be quite different from the explanation for the other", [adding that] "Otto has to look at his notebook while Inga has to look at nothing"
>
> (both quotes from Butler (1998) p. 211)

But from our point of view, Otto's inner processes and the notebook constitute a single, extended cognitive system. Relative to this system, the flow of information is wholly internal and functionally akin to introspection (for more on this, see section 6 following).

One way to try to push the argument is to seek an independent criterion for the perceptual. With this in mind, Martin Davies (personal communication) has suggested that it is revealing that Otto could misread his own notebook. This opening for error may, Davies suggests, make the notebook seem more like a perceived part of the external world than an aspect of the agent. But parity still prevails: Inga may misremember an event not due to an error in her memory store but because of some disturbance during the act of retrieval. The opening for error does not yet establish that the error is, properly speaking, perceptual. It only establishes that it occurs during retrieval.

A slight variant, again suggested by Martin Davies, is that perception (unlike introspection) targets a potentially public domain. Notebooks

and databases are things to which other agents could in principle have access. But (the worry goes) my beliefs are essentially the beliefs to which *I* have a special kind of access, unavailable to others.

There is, of course, something special about Otto's relation to the information in the notebook, in that (as we commented in the original paper) Otto more or less automatically endorses the contents of the notebook. Others, depending on their views of Otto, are less likely to share this perspective. But this is not a special kind of access so much as a special kind of cognitive relationship.

But why suppose that uniqueness of access is anything more than a contingent fact about standard biological recall? If, in the future, science devised a way for you to occasionally tap into my stored memories, would that make them any less *mine*, or part of my cognitive apparatus? Imagine, for that matter, a form of MPD (Multiple Personality Disorder) in which two personalities have equal access to some early childhood memories. Here we have (at least arguably) a case where two distinct persons share access to the same memories. Of course, one may harbor all kinds of reasonable doubts about the proper way to conceptualize MPD in general. But the point is simply that it seems to be at most a contingent fact that I and I alone have a certain kind of access to my own biologically stored memories and beliefs.

Before leaving this topic, I want to briefly mention a very interesting worry raised by Ron Chrisley (personal communication). Chrisley

notes that as a child, we do not begin by experiencing our biological memory as any kind of object or resource. This is because we do not encounter our own memory perceptually. Instead, it is just part of the apparatus through which we relate to (and experience) the world. Might it be this special developmental role that decides what is to count as part of the agent and what is to count as part of the (wider) world?

Certainly, Otto first experiences notebooks (and even his own special notebook) as objects in his world. But I am doubtful that this genuine point of disanalogy can bear the enormous weight that Chrisley's argument requires. First of all, consider the child's own bodily parts. It is quite possible, it seems to me, that these are first experienced (or at least simultaneously experienced) as objects in the child's world. The child sees its own hand. It may even want to grab the toy and be unable to control the hand well enough to do so. The relation here seems relatively 'external', yet the hand is (and is from the start) a proper part of the child.

Perhaps you doubt that there is any moment at which the child's own hand is really experienced (or at any rate conceptualized) as an object for the child? But in that case we can surely imagine future non-biological (putatively cognitive) resources being developmentally incorporated in just the same way. Such resources would be provided so early that they, too, are not first conceptualized as objects (perhaps spectacles are like this for some of us already).

Contrariwise, (as Chrisley himself helpfully points out) we can imagine beings who from a young age are taught to experience even their own *inner* cognitive faculties as objects, courtesy of being plugged into bio-feedback controllers and trained to monitor and control their own alpha rhythms etc.

The developmental point, though interesting, is thus not conceptually crucial. It points only to a complex of contingent facts about human cognition. What counts in the end, though, is the resource's current role in guiding reasoning and behavior, not its historical positioning in a developmental nexus.

6. Perception, Deception and Contested Space

In a most interesting and constructive critique of the extended mind thesis, Kim Sterelny (In Press) worries that Clark and Chalmers underplay the importance of the fact that our epistemic tools (our diaries, filo-faxes, compasses and sextants) operate in a "common and often contested" space. By this, he means a shared space apt for sabotage and deception by other agents. As a result, when we store and retrieve information from this space, we often deploy strategies meant to guard against such deception and subversion. More generally still, the development and functional poise of perceptual systems is, for this very reason, radically different from the

development and functional poise of (biologically) internal routes of information flow. The intrusion of acts of perception into Otto's information retrieval routine thus introduces a new set of concerns that justify us in not treating the notebook (or whatever) as a genuine part of Otto's cognitive economy.

Sterelny does not mean to deny the importance of 'epistemic artefacts' (as he calls them) in turbo-charging human thought and reason. Indeed, he offers a novel and attractive co-evolutionary account in which our ability to use such artifacts both depends on, and further drives, a progressive enrichment of our internal representational capacities. In this way:

> "Our use of epistemic artifacts explains the elaboration of mental representation in our lineage and this elaboration explains our ability to use epistemic artifacts" Sterelny (In Press)

What he does mean to deny, however, is that the use of such artifacts reduces the load on the naked brain, and that the brain and the artifacts can coalesce into a single cognitive system. Instead, he sees increased load and a firm boundary between the biological integrated system and the array of props, tools and storage devices suspended in public space. I tend to differ on both counts, but will here restrict

my comments to the point about the boundary between the agent and the public space.

Within the biological sheath, Sterelny argues, information flow occurs between a "community of co-operative and co-adaptive parts [that are] under selection for reliability" Over both evolutionary and developmental time, the signals within the sheath should become clearer, less noisy, and less and less in need of constant vetting for reliability and veridicality. As soon as you reach the edge of the sheath however, things change dramatically. Perceptual systems may be highly optimized for their jobs. But it is still the case that the signals they deliver have their origins in a public space populated in part by organisms under pressure to hide their presence, to present a false appearance, or to otherwise trick and manipulate the unwary so as to increase their own fitness at the other's expense. Unlike internal monitoring, Sterelny says:

> "…perception operates in an environment of active sabotage by other agents {and] often delivers signals that are noisy, somewhat unreliable and functionally ambiguous" Sterelny (In Press).

One result of all this is that we are forced to develop strategies to safeguard against such deceptions and manipulations. The cat moves gingerly across the lawn and may stop and looks very hard before trusting even the clear appearance of a safe passage to the other side. While at a higher level by far, we may even deploy the tools of folk

logic and consistency-checking (here, Sterelny cites Sperber (forthcoming)).

The point about vulnerability to malicious manipulation is well-taken. Many forms of perceptual input are indeed subject, for that very reason, to much vetting and double-checking. I do not think, however, that we treat all our perceptual inputs in this highly cautious way. Moreover, *as soon as we do not do so*, the issue about extended cognitive systems seems to open up (see below). As a result, I am inclined to think that Sterelny has indeed hit on something important here, but something that may in the end be helpful, rather than harmful, to the extended mind account.

Take the well-known work on magic tricks and so-called "change blindness" (for a review, see Simons and Levin (1997)). In a typical example of such work you might be shown a short film clip in which major alterations to the scene occur whilst you are attending to other matters. Often, these alterations are simply not noticed. Once they are drawn to your attention, however, it seems quite amazing that you ever missed them. The art of the stage magician, it is often remarked, depends on precisely such manipulations. We are, it seems, remarkably vulnerable to certain kinds of deception. But this, I want to suggest, may be grist to the extended mind mill. For the reason we are vulnerable in just those kinds of cases is, I would argue, because we are relying on an ecologically sound strategy of treating the external scene as a stable, reliable substitute for internally-stored memory traces. In short, our brains have decided (if you will allow

such loose talk for a moment) that on a day to day basis the chances of these kinds of espionage are sufficiently low that they may be traded against the efficiency gains of treating the perception-involving loop as if it were an inner, relatively noise-free channel, thus allowing them to use the world as 'external memory' (O 'Regan (1992), O 'Regan and Noe (2001)).

It is important, in our story about Otto, that he too treats the notebook as a typically reliable storage device. He must not feel compelled to check and double-check retrieved information. If this should change (perhaps someone carefully does begin to mess with his external stored knowledge base), and Otto should notice the change and become cautious, the notebook would at that point cease to count as a proper part of his individual cognitive economy. Of course, Otto might wrongly become thus suspicious. This would parallel the case of a person who begins to suspect that aliens are inserting thoughts into their head. In these latter cases, we begin to treat biologically-internal information flow in the cautious way distinctive of perception.

In sum, I think Sterelny is right to pursue this kind of issue. But what emerges is not so much an argument against the extended mind as a way of further justifying our claim that in some contexts signals routed via perceptual systems are treated in the way more typical of internal channels (and vice versa, in the case of standard thought-insertion). To decide, in any given case, whether the channel is acting more like one of perception or more like one of internal information

flow, look (in part) to the larger functional economy of defenses against deception. The lower the defenses, the closer we approximate to an internal flow.

Sterelny might reply to this by shifting the emphasis from the extent to which an agent actually does guard against deception and manipulation to the extent to which they are, as a matter of fact, vulnerable to it. Thus the fact that we are vulnerable to the magician's art may be said to count for more than the fact that in being thus vulnerable we treat (as I tried to argue) the perceptual route as a quasi-internal one. But this seems unprincipled, since given the right 'magician' (say, an alien able to directly affect the flow of energy between my synapses) all routes seem about equally vulnerable. Recall also that false beliefs can (as we noted earlier in this essay) be generated in biological memory by many a good psychologist. Or, for that matter, the many rather bizarre ways in which biological memory and reason can be systematically impaired (for example, the patients whose memories, like their ongoing experience, exhibit hemispatial neglect (Bisiach and Luzzatti (1978), Cooney and Gazzaniga (2003)). What seems (to me) to count is not vulnerability as such but rather something like our 'ecologically normal' level of vulnerability. And our actual practices of defense and vetting are, I claim, rather a good guide to this. If Otto doesn't worry about tricksters copying his writing and adding false entries, maybe that is because the channel is as secure as it needs to be.

There is, finally, a large and I suspect un-resolvable issue still waiting in the wings. For present purposes I am happy to have shown (or tried to show) that the very large differences that Sterelny highlights do not in fact obtain in the kinds of case Clark and Chalmers meant to imagine. But nonetheless I must concede (to Sterelny and to others) that the functional poise of information stored in public space is probably never *quite* the same as that of information stored using our inner biological resources. Might this itself secure the conclusion that information thus stored cannot count towards an agents stock of dispositional beliefs? To do so would require a strong intervening premiss. One such premiss would be, for example, the claim that perfect identity of functional poise is essential if non-biologically stored information is to count. But such a requirement is surely too strong. For all we know, the fine details of functional poise differ from person to person and hour to hour. This point is merely dramatized by those alien beings whose recall is (let's imagine) not subject to hemispatial neglect, cross-talk or error: do these differences make a difference? Is the alien whose recall is fractionally slower than ours, or fractionally faster, or much less prone to loss and damage, to be banned from the ranks of true believers? To demand identity of functional poise is surely to demand too much.

But just 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 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.

7. An Alternative Ending?

Recall Adams and Aizawa's worry that the inner/outer elements form at best a motley, not the kind of causally unified set needed to support a real science, and their insistence that "the cognitive must be discriminated on the basis of underlying causal processes" (op cit p. 52) . In reply (section 3 above) I mooted that there might be great variety among the inner, and paradigmatically cognitive, elements themselves: fully as much variation, perhaps, as between the inner and outer. This raises, however, the possibility of an alternative

reading of the Clark and Chalmers argument itself. Perhaps the real moral of the story is that the realm of the mental is itself too dis-unified to count as a scientific kind?

This idea was first suggested to me by Jesse Prinz and was to be investigated in a joint project (Clark and Prinz, stalled). The claim of that paper was to be that:

> "there is no unified, coherent understanding of the very *idea* of 'mind' at work in various philosophical and scientific projects all of which claim to be studying aspects of the mental…"
>
> and that:
>
> " not only is there no satisfying definition available, there is not even a useful shared scientific understanding, guiding prototype, or loosely connected web of salient properties and features. ..there are no signs that we are here dealing with any natural kind. …nor…with anything perhaps more nebulous, but nonetheless capable of legitimating the mind as a proper object of scientific study"
>
> Both quotes from Clark and Prinz, (Stalled).

Evidence for this rather dramatic claim could be found, we suggested, in the endless philosophical debates over the applicability of mental predicates to an incredibly wide variety of cases, such as:

thermostats (Dennett (1987)), paramycia (Fodor (1986)), language-less animals (McDowell (1994)), swampmen, computers (Searle (1980)), sub-personal 'cognitive' activity in general(Searle 1992). Not to mention non-human animals, fetuses, pre-linguistic infants, coma patients and now, of course, *extended cognitive systems* such as Otto and his trusty notebook. The point we wanted to make was that there was no easy consensus among 'suitably trained observers' concerning the distribution of minds and mentality in nature and artifice. We just don't know a mind when we see one. Could the reason for this be that there simply aren't any there? Might the Extended Mind debate form part of a reductio of the very notion of Mind in Cognitive Science?

In response to this suggestion, I would concede that the notion of 'mind' as it is now used is torn between its roots in the idea of conscious experience and occurrent thoughts, and its extension into the realm of non-conscious processes and long-term stored knowledge. It is this latter extension that opens the door to the Extended Mind argument. One good way of reading that argument, I have long thought, is as a demonstration that if you allow non-conscious processes to count as properly mental, the physical basis of the mental cannot cannot remain bound by the ancient barriers of skin and skull. Nor should it be thus bound since, (as argued in section 4), attempted defenses that stress occurrent processes (there, of ultimate control and choice) will surely shrink Mind too small, ruling out much that we want to count as mental and cognitive even inside the head. But since for many tastes, the Extended Mind story

bloats Mind too large, could we not conclude that the idea of the mental is terminally unstable? Couldn't we just *eliminate the mind*?

I don't think so (hence the perhaps-permanently-stalled status of the Clark and Prinz paper). For as I noted in section 3, despite the mechanistic motley, we may still aspire to a science of the mind. Granted, this will be 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. It will be a science that needs to cover a wide variety of mechanistic bases, reaching out to biological brains, and to the wider social and technological milieus that (I claim) participate in thought and reason. It will *have* to be that accommodating, since that very mix is what is most characteristic of us as a thinking species (see Clark (2003)). If we are lucky, there will be a few key laws and regularities to be defined even over such unruly coalitions. But there need not be. The science of the mind, in short, won't be as unified as physics. But what is?

In sum, I am not ready to give up on the idea of minds, mentality and cognition any day soon. The Extended Mind argument stands not as a reductio but as originally conceived: a demonstration of the bio-technological open-ness of the very ideas of mind and reason.

Conclusions.

The notion of the Extended Mind draws strong reactions. Many feel it is patently false. These same people tend to feel that the mind is simply and obviously just the activity of the brain. Others regard it as patently true, and they tend to be those who identify the mind with an essentially socially and environmentally embedded principle of informed agency (ie the fans of situated cognition). My own feeling is that we have not yet reached the philosophical or scientific bottom of this debate. There is something important to be said, for example, about the role of emotion in constantly coloring and informing cognition, and something (perhaps along the lines of Damasio REF) about the way our ongoing sensing of our own biological body-state informs our sense of self. There is much to be said about the way our sense of what we know is, at bottom, a sense of what kinds of information we can easily and reliably exploit in the pursuit of our daily goals and projects (for a detailed meditation on this theme, see Clark (2003)). The critical role of conscious awareness and occurrent thought in the overall debate over what is mental and what is not is worrisomely unclear, and will probably remain so until we have a better understanding of the neural roots of qualitative experience. Finally, the consistent (though to my mind unattractive) option of simply restricting the realm of the mental to that of occurrent conscious processing probably bears further thought and investigation, though not, I expect, by me.

So does Leonard (the protagonist of Memento) really increase his stock of beliefs every time he gets a new body tattoo? Better wait for the sequel.

References

Adams, F and Aizawa, K (2001) "The Bounds of Cognition" *Philosophical Psychology* 14:1 p.43-64

Bisiach, E and Luzzatti, C (1978) "Unilateral neglect of representation spce" *Cortex* 14 129-133

Butler, K. (1998). *Internal Affairs: A Critique Of Externalism In The Philosophy Of Mind.* Dordrecht, Kluwer.

Churchland, P. S. and Sejnowski, T. (1992). *The Computational Brain*. Cambridge, MA: MIT Press.

Clark, A (2003) *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence* (Oxford University Press)

Clark, A. And Chalmers, D. (1998). "The Extended Mind." *Analysis* 58: 10-23
Reprinted in (P. Grim, ed) *The Philosopher's Annual*, vol XXI, 1998.

Clark, A and Prinz , J (Stalled) "The Absence of Mind" (ms)

Cooney, J and Gazzaniga, M (2003) "Neurological disorders and the structure of human consciousness" *Trends in Cognitive Sciences* 7:4:161`-165

Decety, J. and Grezes, J. 1999 "Neural Mechanisms Subserving the Perception of Human Actions". *Trends in Cognitive Sciences,* 3 :5: 172-178.

Dennett, D (2003) *Freedom Evolves*

Graham-Rowe, D (1998) "Think and It's done" *New Scientist,* October 17, 1998.

Norman, D (1999) *The Invisible Computer* (MIT Press, Cambridge, Ma)

O'Regan, J.K. (1992) "Solving the "real" mysteries of visual perception: the world as an outside memory" *Canadian Journal Of Psychology* 46:3:461-488

O'Regan, J.K., and Noe, A In Press, 2001 "A Sensorimotor Account of Vision and Visual Consciousness" *Behavioral and Brain Sciences* 24:5.

Simons, D and Levin, D (1997) "Change Blindness" *Trends in Cognitive Sciences,* 1: 7: 261-267

Sperber, D (forthcoming) "An evolutionary perspective on testimony and argumentation" *Philosophical Topics*

Sterelny, K (In Press) "Externalism, Epistemic Artefacts and the Extended Mind" R Schantz (ed) *The Externalist Challenge: New Studies on Cognition and Intentionality* (de Gruyter, Berlin and NY)