Leonard's System: Why Doesn't It Work?
Joseph Levine, UMass Amherst

Near the beginning of Memento, Leonard argues forcefully that his disability, his inability to lay down new memories, isn't really all that disabling after all. He claims that memory is overrated, and that his system of writing everything down and constantly checking his notes provides him with sufficient information to get by successfully. He contrasts most peoples' reliance on memory with his reliance on "the facts"; that is, what he's written down. Given the research that shows memory to be much more a process of construction than of pure retrieval,[1] and given the relative permanence of what is written down, it might seem offhand that he has a point. Most of us don't bother to write down most of what we want to remember, since, after all, we can remember. But maybe Leonard is right. Perhaps if we were to rely more on what we could write down, and less on memory, our epistemic situation - the level of justification that our current beliefs achieve at any given time, as well as the range of facts with which we would be acquainted - would improve. Or, at least, it wouldn't be much, if at all, worse off.

As anyone who has seen the movie knows, the plot is presented backwards, with each new scene depicting events that occurred before those of the preceding scene. While this technique has been used before, it is especially interesting in this case since it puts us, in a way, in the same epistemic situation as Leonard. During any scene we have as little to go on as he has. All we know is what is right there in the moment, including the notes about past events that are available. Leonard can't remember, and we're in the same boat since we haven't seen the past yet.

---

[1]For a good overview of the research on memory, see Reisberg (2006), Part III.

When Leonard makes his little speech about how he doesn't need memory to get by, that he does just fine with his notes and his "facts", the movie encourages us to take what he says seriously. At that point we've seen him kill someone who seemed to be a threat to him - according to his notes, anyway - and in general he appears to be functioning fairly well. So maybe we're not really going to buy this "memory is overrated" argument - it *is* just a movie, after all - but the picture presented has a definite air of plausibility about it.

Of course, by the end of the movie it becomes all too clear that, far from functioning fairly well with his notes and his "facts", Leonard is totally clueless. As we learn the course of events that led up to the initial scenes in which he seemed to be doing okay, we realize that his epistemic situation in those early (that is, temporally later) scenes is so weak as to render him powerless to control his life in any meaningful way. The notes that were supposed to provide his link to the past serve to consistently mislead him. None of his inferences about his present situation or how it came about are reliable. The world is worse than a buzz of confusion to him, since he thinks he knows what's going on while suffering massive delusion. For Leonard, "everything he knows is wrong".

So now the question presents itself: what was wrong with his method? Why is memory so crucial to maintaining an epistemic position that allows one to function, even minimally, in the world? Why don't his notes and strategies for getting by succeed? What epistemological lessons can we learn from Memento concerning the crucial role that memory plays in providing a cognitive subject with a belief system that adequately

mirrors the world around her?

In this essay I want to explore three approaches to answering these questions: I call them the "qualitative-difference" approach, the "mere-quantitative-difference" approach, and the "architectural" approach. I will begin by discussing the first two approaches, showing why I think they do not adequately explain Leonard's problem. I will then describe the third approach, which incorporates some features of the first two, but in a way that properly explains Leonard's epistemic disability. In the course of defending the architectural approach I will discuss the bearing of Leonard's case on an extremely provocative view in the philosophy of mind, championed by Clark and Chalmers (1998), called the "extended mind" hypothesis. On this view it is wrong to think of the skull, or the nervous system, as a principled boundary between the mind and the outer world. On the contrary, information that we write down, store in our computers, or just possess in our library, is a part of our minds in the same way as what is stored "inside". This view of the mind should find nothing in principle wrong with Leonard's system, so an advocate of the extended mind hypothesis is likely to adopt the mere-quantitative-difference approach to explaining Leonard's epistemic deficit. My argument against that approach will also constitute an argument against the extended mind hypothesis itself.

THE QUALITATIVE- AND MERE-QUANTITATIVE DIFFERENCE APPROACES

On the qualitative-difference approach, the problem Leonard faces is, as the name suggests, a qualitative one; that is, in principle, nothing that is written down as a mere historical record can play the epistemic role that memory plays. By a "mere

historical record" I mean to distinguish cases like Leonard's from more mundane cases where people write things down - like shopping lists - in order to aid or jog their memories. When I consult a shopping list and see that it says "milk", I don't just then learn that I need milk, but rather remember that I do. What's written down functions to prompt a memory experience. With Leonard, however, the notes he writes for himself are not prompts or aids of this sort. They don't remind him of anything. Rather, he is in the position of an historian discovering a document from the past and attempting to figure out what it reveals about events at the time. On the qualitative-difference approach, then, though historical records are of course a valuable source of data about past events, they cannot substitute for memory. Memory provides a different kind of support for our beliefs about the world, without which genuine knowledge (except of the most trivial kind concerning current experience) is not possible.

On the mere-quantitative-difference approach, on the other hand, there is nothing principled about Leonard's weak epistemic position. Historical records don't differ in kind from memories. The problem Leonard faces is purely quantitative in nature. Simply put, he just can't write enough down. If somehow he could write faster, or speak fast enough into a tape recorder (or tattoo himself faster - ouch!), he could in principle use his system to overcome the disabling effects of his memory loss on his epistemic relation to the world. However, as a matter of fact, it isn't possible to write or talk fast enough, and too much data gets lost. The difference between his epistemic position and our own is like the difference between an historian who studies a recent period of Western history, with tons of archival material available to her, and one who studies an

ancient period from which only fragments survive. They are both engaged in the same kind of enterprise, but one just has so much more to go on than the other.

Let's begin now to consider the qualitative-difference approach. What difference in kind might there be between memory and other forms of information storage that might bear on their epistemic status? Tyler Burge's (1995) idea of "content preservation", which he applies both to memory and testimony, might provide the basis for making just such a principled, qualitative distinction. So let's take a digression from Leonard's predicament just to get clear on Burge's view. Afterward, we'll see if it helps to explain why Leonard is so epistemically disabled by his memory loss.

Burge's discussion starts from consideration of a puzzle. But first a few preliminaries. Traditionally, philosophers distinguish "a priori knowledge" from "a posteriori knowledge". The former include mathematics, logic, and purely definitional propositions, such as that all bachelors are unmarried. The latter include most of what we know, from where our keys are to highly theoretical claims in science. While there are many different definitions of the terms "a priori" and "a posteriori" around, let's follow Burge in using the terms in the following manner. One possesses an "a priori warrant" for one's belief, if no perception of a specific situation plays a role in providing the justification for the belief. One possesses an "a posteriori warrant" just in case an experience of a specific situation does provide part of the justification. So, for instance, if my basis for believing that there's a computer screen in front of me right now as I type is that I see it, then my belief that there is a computer screen in front of me right now possesses an a posteriori warrant. However, if my basis for believing that the

Pythagorean theorem holds is my seeing a proof, this particular written expression of the proof is not really part of what justifies my belief, even though of course I accessed the proof through seeing it. Rather, once I understand it, it's my cognitively appreciating the relation between the premises and the conclusion that does the work. Hence, though vision plays a causally necessary role in this case, my warrant for believing the theorem is still a priori.

Now comes the puzzle. Suppose one comes to believe a mathematical theorem on the basis of a fairly long proof. Perhaps the proof itself is, say, 50 lines long, and the conclusion is derived by a valid inference rule from lines 49 and 25. Granted, if the rule used to derive line 50 from the other two is logically valid, then one's warrant for believing the proposition expressed by line 50 is a priori so long as one's warrant for believing the immediate premises, the propositions expressed on lines 49 and 25, is also a priori.

But now consider one's warrant for believing line 25. True, at the time one derived it, one could see how it followed validly from earlier premises (which, we'll assume, were warranted a priori). But that information is now gone. One remembers deriving it validly, and now remembers the result, line 25, and so one uses it to derive the conclusion, line 50. The puzzle is, does this reliance on memory to furnish the premise for the last step render one's warrant for the conclusion a posteriori? If so, very little of what we normally consider a priori knowledge will survive.

Why might the reliance on memory undercut the a priori status of one's warrant for belief in the conclusion of the proof, the theorem? The idea is that memory, like

perception, is an experience. Just as an answer to the question, “How do you know there’s a computer screen in front of you?” is “I see it” - that is, I’m having a particular experience of the computer screen - so too an answer to “How do you know that the proposition expressed by line 25 is true?” might be “I remember proving it” - that is, I’m having a particular experience of a previous mental state. In both cases, one might argue, I am currently relying on an experience of some particular situation or event - what’s outside me in one case and what happened earlier in my mind in the other - to support my current belief. If the one case counts as a posteriori, goes the argument, so too should the other.

Burge argues, however, that the role of memory in the proof case is quite different from the role of vision in the computer screen case. In the latter case, vision is supplying new information, adding to whatever justificatory force might or might not have previously attached to the proposition that there is a computer screen in front of me. In this sense it is correct to say that my belief that there is a computer screen in front of me is epistemically based on the visual experience; it relies on the visual experience for its justification, or warrant. However, in the proof case, where memory, as it were, “delivers” the proposition from line 25 to my current state of mind, to be used in the derivation of line 50, memory isn’t adding a new content, or providing justificatory force; rather, the role of memory in this case is purely “preservative” (hence the name of his article, “Content Preservation”). The idea is that memory functions to preserve within my belief store information that might be needed for later reasoning, and the preservative process of memory functions not only to maintain the propositional content

itself, but also its level and kind of warrant. Memory isn't adding to, or subtracting from the warrant that attaches to the proposition. It is preserving it from its earlier presence in one's conscious awareness, and transferring it and presenting it to one's current conscious awareness. Of course one relies on the mechanism of memory when using the earlier line in a current bit of reasoning, but, for Burge, that is no different from the way we use what's written down on the page to follow the proof. In neither case is the experience itself, of seeing the proof or remembering the line, itself explicitly part of the justification for the conclusion. We might put it this way: What is remembered plays a role, but not the remembering of it itself.

Let's get back to Leonard and his predicament. Suppose Burge is right and memory "preserves content" in the way he suggests. How might this make a difference to Leonard? Well, what we are imagining is the situation in which a piece of information that would have been available to Leonard via memory, had he not suffered the trauma to his brain, is now available in written form. Since he has to perceive the note - whether on a piece of paper or on his body - Burge's notion of content preservation doesn't apply. The information isn't immediately available to him, but rather must be perceived and interpreted. Perhaps this then is the crucial difference between a normal person's ability to rely on memory to make her way in the world and Leonard's need to rely on a written record. For the normal person who remembers what she's seen and heard, the information so acquired is "preserved" by memory; whereas for Leonard, relying on his written records, it must be accessed again through perception and interpretation.

Of course Burge's account of the way memory "preserves content" is quite controversial, but even if we accept it, I don't see how it helps to explain Leonard's problem. What Burge's account can explain is why Leonard's beliefs about the past, and what's going on in the present to the extent this is conditioned by the past, perhaps have a different epistemic status than those of a person with intact memory. Perhaps Leonard couldn't know a theorem that took many steps to prove in the way a normal person could. Perhaps every belief about the past is burdened with an extra layer of required justification, since each bit of information must be perceived and interpreted. But all of this may be true even though the information recorded in his notes is every bit as reliable as that which others get by memory, or what he himself would have obtained by memory had he not suffered from his condition.

One issue is the kind of justification or warrant a belief has - is it a priori or a posteriori, is it "preserved" by memory or based on current perception? This might matter for certain issues, such as the status of mathematical knowledge, as Burge's discussion suggests. But so long as the information is relatively reliable, what kind of justification it has shouldn't matter for the prospects for success in acting on it. For action, what matters is reliability. True, Leonard's information, given its written form, must first be perceived and interpreted before he can make use of it, and these added steps do also add some degree of risk of getting things wrong. But if this were the only problem it wouldn't seem to be so debilitating as his condition clearly is. After all, the perceptual and interpretive mechanisms he employs - his visual and language systems - are tremendously reliable. The added epistemic risk that attends the need to perceive

and interpret what would otherwise have been remembered seems marginal. And let's not forget, as Leonard reminds us, that memory mechanisms themselves are not always tremendously reliable. After all, that is why we make shopping lists for ourselves. I conclude that Leonard's problem cannot be explained by the different epistemic status possessed by what's remembered from what's perceived (even if, as Burge argues, there were such a difference).

What's the alternative explanation? Well, the mere-quantitative-difference approach pins Leonard's deficit totally on the quantity of information he misses by losing his short-term memory. As I put it above, we can think of Leonard as an historian of the ancient world, with only fragments to go on. There is, after all, just so much he can write down, and that amount is so much less than what a normal person can remember, it's no wonder he finds himself at such an epistemic disadvantage. Perhaps that's all there is to it.

While I think it's right to say that the small quantity of information he has available is crucial to explaining Leonard's epistemic deficit, we need to put a little more emphasis on the term "available" here to get a full explanation of what's going on. True, he can only write down so much. But even if he could write down everything a normal person would remember, he would still have a problem. The problem is that it doesn't do him any good sitting on paper. It has to get inside his mind, where his memories would have been, to be of any use to him. So while the qualitative-difference idea concerning the special "preservative" status of memory proved irrelevant to explaining his deficit, the fact that what's written down has to be perceived and interpreted before it can be of

use is quite relevant. In this sense I want to claim that both the qualitative-difference and the mere-qualitative-difference approaches possessed a grain of truth. As the qualitative-difference approach emphasizes, the fact that Leonard must perceive and interpret what's written down - steps not necessary for what's remembered - is crucial to explaining his epistemic deficit. On the other hand, as the mere-quantitative-difference approach insists, the issue in the end does come down to a matter of how much information is available to Leonard. The way to incorporate both of these insights is through the architectural approach, to which I now turn.

THE ARCHITECTURAL APPROACH

The architectural approach is based on the idea that the mind is a kind of information processing device with a computational architecture, as is the case with your computer. Computers contain central processors, working memory, hard drives, input devices (keyboards, microphones, cameras), and output devices (monitors, printers, speakers). According to the currently popular "computational model", minds also possess an architecture, which it makes sense to investigate empirically. While there are a number of issues here that fall under the heading of "cognitive, or computational architecture", the one that I want to exploit for the purpose of explaining Leonard's disability is the architectural distinction between cognitive "modules" and "central processes". Let me explain this distinction, and then show how it helps to understand Leonard's predicament.

Jerry Fodor (1983) divides mental operations into two basic kinds: those that take place within input systems, which he calls "modules", and those that take place within

what he calls “central systems”, which are not modular. To oversimplify, assume that we have six input systems; one each for the five senses, and one for interpreting and producing language. According to Fodor, each of these systems is modular, in the sense that it works largely in isolation from the other input systems, as well as the central systems. The idea is that an input module, say vision, takes stimulations of nerve endings as input and delivers a description of the relevant part of the environment as output. One of the crucial features of a modular processor is that it doesn’t have access to information that is available to the mind as a whole, but only to a restricted range of information that is stored within it.

One way to see what Fodor means here is to consider a visual illusion. You look at a stick partially submerged in water and it looks bent. After lifting it out of the water you see that it’s really straight, it’s just that the refraction of the light as it travels through the water made it look bent. But now, when you stick it back in the water, it still looks bent, even though you know perfectly well that it’s straight. What’s going on? Well, according to Fodor, the persistence of the illusion that the stick is bent reflects the fact that your visual system is modular. In computing the shape of the stick from the light hitting the retina, the visual system doesn’t know about the refraction of light in water. What’s more, your visual system also doesn’t have access to the information that you, as a matter of what’s stored in your central system, have available; namely, that it’s a straight stick. The point is that seeing isn’t believing after all. What you see is determined by what your visual system figures out, but what you end up believing on the basis of what you see takes account of everything else you know as well. In this

case, you will believe the stick is straight even though it looks bent. You know, as it were, much more than your visual system does about this.

This distinction between how things look to you in the stick-in-the-water case and what you actually believe about the shape of the stick highlights the particular way in which modular processing differs from central processing, the non-modular system in the mind within which all-things-considered belief is determined. We can put it this way: modular processing is “local” in character, whereas central (non-modular - I’ll let this be understood from now on) processing is “global”. Taking the visual system as our example again, the information processing taking place within it applies to two sources of data for delivering verdicts about the spatial layout of objects in the subject’s environment: the light hitting the retina, and whatever information is stored within the visual module. In deciding how to visually represent the shape of the stick - i.e. present how the stick looks - the visual system only consults these two local sources of data. What you know about its shape from having examined it before isn’t part of its data base.

On the other hand, central processing, the kind that determines what you actually believe, seems to be global in character. When trying to figure out what to believe about a particular situation, almost anything you believe about any other topic - whether it be general knowledge, memories of particular events, or current experiences - is potentially relevant. What’s more, when deliberating between two, mutually exclusive possibilities for what to believe, a large part of what decides between the two hypotheses are certain global features of the overall belief system one would have were

one to adopt one or the other to believe.

Perfectly mundane matters as well as highly theoretical ones exemplify both points. For example, suppose you are driving on the highway and you see a sea of brake lights in front of you all of a sudden. What you see is just that - a lot of red lights coming on in your visual field. But what you come to believe depends on all sorts of other things you might happen to believe at the time. If you believe you are on your way into Washington, D.C., and you also believe that an Al Qaeda video has just been released, you might come to believe that there is a police road block ahead, checking for weapons. A belief about something that happened half way across the world was brought to bear on how to interpret the import of red lights flashing in your visual field.[2]

As for how global features of entire belief systems affect which particular belief we adopt in a particular situation, consider how most of us react to reports of supernatural events, even if we ourselves perceived the allegedly supernatural event. We refuse to believe it. Why? Well one way of explaining our reluctance to believe that a genuinely supernatural event has occurred is that doing so would conflict with deeply held common sense and scientific beliefs. But of course we could remove the conflict by making an ad hoc adjustment in our overall belief system, perhaps by believing that normal laws of nature were suspended in this instance. While this may remove the formal contradiction between our general common sense and scientific beliefs on the one hand and the belief that a supernatural event has just occurred on the other, the resulting belief system as a whole becomes clearly less simple and elegant than the

[2]For a really nice example of this idea of potential relevance from anywhere as it applies

alternative; namely, that the report was false, or we ourselves suffered from an illusion. Features such as generality, simplicity, and elegance, which clearly play a large (though not very well understood) role in how we decide what to believe, are global features of whole belief systems, not merely of individual beliefs or the relation between small sets of data and conclusions drawn from them.[3]

It's important to note that there are two sides to the global character of central processing, both of which bear on Leonard's situation. First, independently of how we in fact reason, canons of good reasoning demand that we take account of both the potential relevance of almost any belief to any other and the global features of entire belief systems. By saying that this is a rational demand, what I mean, in the first instance, is that it reasoning in accordance with this demand yields beliefs that are much more likely to be true than reasoning in violation of it. Processes that determine beliefs according to these global principles are highly reliable.

Secondly, not only is following such a globally sensitive procedure a rational demand, but we seem in fact, to a very remarkable extent, to follow such procedures. That is, the kind of information processing that goes on in our central system, the kind that determines what we believe all-things-considered, does seem sensitive to these global features of our belief system. Our thinking seems responsive both to the relevance (and irrelevance) of information from widely disparate areas and to the relative simplicity of vast bodies of beliefs. How we are able to do this is still not well

---

in science, see Antony (2003).

[3]For a classic discussion of this issue, see Quine and Ullian (1978).

understood at all.[4] Nevertheless, we do do it, and given the way the world is put together, our ability to do it seems to be crucial to our tremendous success in arriving at largely true beliefs about the world.

There is one more element that needs to be added before the architectural approach to explaining Leonard's epistemic problem can be presented; and that is the notion of the "domain of an operation", or, equivalently, the "set of representations over which an operation is defined". An operation in an information processing system applies to a representation (or a set of representations), taking it (or them) as input to the operation, and yielding a representation (or set of representations) as output. If one reasons, say, from the belief that the stick looks to be bent to the belief that the stick is bent, we can think of this as an operation that took the first belief - a representation of the way the stick looks - as input, and delivered the second belief - a representation of how the stick is - as output. The domain of an operation, then, is just the set of representations that that operation is capable of accepting as inputs.

With this idea in mind, one way to characterize the function of an input system is this: input systems serve to transform information from a form in which it is not in the domain of the operations that make up central processes into a form in which it is.[5] So again, consider a simple case of looking at a stick in water. The world contains this

---

[4]Fodor (1983 and 2000) argues, in fact, that we haven't a clue how this is done, and that the problem threatens the viability of a computational model of such processes. Others argue the problem is not nearly so intractable. See Ford & Pylyshyn (1996) for several papers that take opposing views on this question.

[5]This leaves out an important aspect of what input systems do, which is to figure out what information is in fact out there. That they don't do this perfectly is why there are illusions. But for present purposes this can be ignored.

information, that the stick is submerged in the water (and that it is straight). Suppose I want to know what's going on with the stick in question. This information is out there in the world - the straight stick is right there in the water - but the central processes responsible for determining what to believe about the stick and its shape can't access this information; their operations don't work on information in that form, at that location. (Rather, they work on representations in the relevant portion of the brain.) But when I look at it, my visual system takes that information[6] and transforms it into a form, and puts it into a location, which make it accessible to the central processes whose job it is to figure out what's going on with the stick.

These features we've been discussing - the modularity of input systems, the global nature of central operations, and the constraints on which representations are included in the domain of the relevant operations - all fit under the rubric of cognitive, or computational architecture. So the architectural approach to explaining Leonard's disability is the approach that explains it in terms of these features. Let's see how this works.

APPLYING THE ARCHITECTURAL APPROACH

I said above that two elements from the first two approaches would be employed in this approach: the fact that Leonard has to perceive and interpret the information that he's written down (along with the snapshots he's taken, of course) and the fact that he has access to much less information than he would have had had his memory been

[6]Again, this isn't quite right. What the visual system does is take the information concerning the light hitting the retina, which, in the normal situation, preserves the information concerning the shape and location of the stick.

intact. To see how the architectural approach works, incorporating both of these elements within it, let's start by imagining a somewhat different scenario. Suppose that Leonard - contrary to what is in fact feasible - is able to write down every bit of information he might have stored in memory had he not suffered from his disability. To draw out the contrast I want to focus on, imagine as well that Leonard has a duplicate, Leonard*, who is just like Leonard in every respect except that his memory systems are all intact. So for every item that Leonard* learns (by perception, testimony, reading, or whatever) and remembers, Leonard learns the same item, but instead of remembering it he writes it down. Thus, if we include all the information that Leonard has written down in his "data base", we can say that he doesn't suffer from a quantitative deficit of information when compared to Leonard*.

Now let's consider a situation in which both Leonard (and Leonard*) has to decide what to believe. Suppose he has to decide whether or not to trust what someone is saying to him. For instance, consider the scene in which Natalie, after leaving her apartment, returns and tells Leonard that her boyfriend beat her. For Leonard*, who can remember Natalie and his history with her, it would be obvious right away that he can't trust what she says. For him, the task of deciding whether or not to trust what's being told him is the usual one we all face. Most of us can come up with a pretty reasonable idea whether to trust what's told us most of the time.

Now let's look at Leonard's situation. Remember, everything in Leonard*'s memory that provides him with evidence concerning Natalie's trustworthiness is available in written from to Leonard. So in one sense he has all the evidence he needs.

Leonard* employs that evidence by applying whatever central operations are relevant to such a decision directly upon the evidence. His memory store contains representations all of which (or enough of which) are included in the domain of those operations. But how is Leonard to use his information? What's written down on paper, no matter how closely he places it to his head, is not within the domain of the operations that figure out what to believe. Hence that information is useless unless it is first transformed into a form, and into a location (i.e. in his brain) where the relevant operations can make use of it. To do that he needs to see it and interpret it. This is precisely the problem.

Quantity of information per se, given our fiction that he can write down everything relevant, is not the problem. Instead, the problem is the inaccessibility of that information; or, to put it another way, the format in which the information is stored. Since Leonard has to perceive and interpret each individual piece of information - all of which is already in the right form and location for Leonard* - he has two severe limitations on his ability to use it: (1) there is a channel limitation on how much information can be processed through his perceptual input system at a time, and (2) when something gets in there it doesn't stay very long - after all, the mechanism that stores information in long-term memory is broken. So Leonard never is in the position of being able to utilize all the information, or evidence, that bears on the question at hand at one time. Thus what his central operations are so good at, detecting these global features of extremely large sets of data, never get the chance to be employed. During the course of any central reasoning operation Leonard is always limited to the

information that has most recently been perceived.[7]

Let me sum up the main points of the architectural approach. Figuring out what to believe involves central processes that somehow operate on large stores of information at once, information that is stored in a designated location and represented in the appropriate form. Memory is precisely that store of information. Since Leonard must perceive and interpret - that is, process through an input system (actually two, vision and language comprehension) - each piece of information he's stored on paper or on his person, his central processes never get the opportunity to operate on all of it at once. Thus even if he can write fast enough, he suffers from insufficient quantity of information, because of the necessity of perceiving and interpreting what he's written. In this way, as I claimed earlier, the architectural approach does incorporate elements from the first two approaches.

THE EXTENDED-MIND HYPOTHESIS

I want now to turn to the bearing of Leonard's case on the "extended mind" hypothesis. As I said earlier, I think that for an advocate of that hypothesis, only the mere-quantitative-difference approach really makes sense, and the limitations of that approach for explaining Leonard's problem show what's wrong with the extended-mind hypothesis itself. I will first briefly present the view in a little more detail, and then explain how it conflicts with the architectural approach to Leonard's problem defended

[7]Of course he also has available, within the domain of the central operations, all of the information stored in memory from before his trauma. Indeed, without that, he would be altogether unable to make it through the day. But the point is that so much of the information upon which his most recent decisions need to be based postdate the trauma, and so is usually unavailable to the relevant operation.

in this paper.

In an influential paper, Clark and Chalmers (1998) defend the claim that the mind isn't confined to what's enclosed within the skull, or even the central nervous system. Rather, instead of there being a principled division between the mind and the world, better to think of the mind as spreading itself out into the world, including within it information that is written down, stored on computer disks and hard drives, and, perhaps, even what's contained on the internet. While clearly for practical purposes we tend to distinguish what's "inside" from what's "outside", their idea is that this isn't a theoretically principled division. The information that I can read on a piece of paper is just as much a part of my mind as what's stored in memory.

To make their case they describe an example involving two people, Inga and Otto, both of whom form the intention to go to a certain location to see a museum exhibit. However Inga remembers where the museum is, so once she forms the intention to see the exhibit she then relatively immediately forms the intention to go to the relevant location. Otto, on the other hand, is suffering from Alzheimer's, so he has to write everything down in a notebook which he keeps on his person and consults constantly. So when he forms the intention to see the exhibit, he first looks up the address in the notebook, and then forms the intention to go to that location. Their claim is that given Otto's mode of interaction with the information in his notebook, he should count as believing that the museum is at the address in question even before looking at the notebook, just by virtue of his having the information stored there. In terms of what Inga and Otto believe, Clark and Chalmers maintain, there isn't a principled difference

between them. Except for details that don't matter for the question at issue, their case of Inga and Otto is strikingly similar to our case of Leonard and Leonard*.

The extended-mind hypothesis might seem to be a natural extension of "semantic externalism", a doctrine that has many adherents in both philosophy of mind and philosophy of language. Semantic externalism is the view that the contents of many of our cognitive states - our beliefs, desires, thoughts, etc. - are partly determined by the world around us. To get the idea, consider Hilary Putnam's famous thought experiment involving "Twin Earth" (Putnam 1975). Putnam asks us to consider a world just like ours, at least superficially, that he calls "Twin Earth". On Twin Earth there is a substance that looks, tastes, and behaves just like water does on Earth, except it's made of a totally different chemical compound, "XYZ" for short. Here's the question: when I think, here on Earth, that water is liquid at room temperature, is my thought about our water only, H_2O, or also about Twin-water, XYZ? Putnam argues, convincingly, that my thought is only about H_2O. That is, what my thought is about, what determines its content, is partly a matter of what stuff I interact with. In this sense, the external world plays a role in determining the content of my thoughts.

Suppose one accepts this view (not everyone does, of course,[8] but, as I mentioned above, it's widespread among philosophers of mind and language). If what's outside my mind plays a role in determining what my thoughts mean, then is it such a stretch to say that my thoughts <u>themselves</u> literally extend out into the world beyond my skull? If the nature of water, that it's H_2O, partly constitutes what I'm thinking about

[8]For a particularly well developed critique of externalism, see Segal (2000)

when I think about water, then why couldn't what's written on a piece of paper, or tattooed on my arm, not count as part of my mind?

In fact I think there is a world of difference between the doctrine of semantic externalism and the extended-mind hypothesis.[9] While the former is quite plausible (though still controversial), it lends no support that I can see to the latter. Furthermore, our reflections on Leonard's disability I think show what's wrong with the extended mind hypothesis. Let me take these points in order.

First, externalism about the contents or meanings of thoughts embodies the following idea. Many of an object's properties or features are determined not only by what it's like in itself, intrinsically, but also by how it is related to other objects. So, to borrow an example from Antony (1995), while Michael Jordan is over six and a half feet tall in his own right, he is (or was) a basketball player by virtue not only of his intrinsic features, but also by virtue of there being an institutional setting, involving thousands of others, within which he can play basketball. In this sense one is an externalist about the feature of being a basketball player.

However, despite the fact that without thousands of others related to him in a complex way - that way that constitutes the institution of basketball - Michael Jordan wouldn't be a basketball player, it is still clear where the boundaries of being a basketball player are; the boundaries of Michael Jordan himself of course. Basketball players are individual, discrete entities, with limits that coincide with the limits of

[9]Clark and Chalmers themselves note the difference, so they don't argue that the extended mind hypothesis just follows from semantic externalism. They do see it as an extension of that doctrine, though, albeit a radical one.

individual human bodies. There is no “extended basketball player” hypothesis that follows from being an externalist about the feature of being a basketball player. So too, just because which thought a mind has is determined partly by facts outside the skull, doesn’t mean that the mind itself extends beyond the skull.

The foregoing was meant to show that adopting externalism about the contents of mental states does not in fact lead naturally to the extended mind hypothesis. But from our discussion above concerning Leonard, we see specific reasons for rejecting that hypothesis. First of all, to the extent that the distinction between an input system and a central system is principled, that enforces a distinction between the “inside” and “outside” that the extended mind hypothesis rejects. After all, what is an input system, if not a mechanism for getting what is outside the mind inside the mind? Calling something an “input” entails there is a boundary to cross.

What then determines the boundary for the “inside”, the mind’s central system? The natural answer is this: the central system includes the mechanisms that embody its principal operations together with the objects upon which those processes operate. Central system operations are those globally sensitive computational operations that determine what we believe, and the only objects on which they operate are those that have been properly processed through one of the input systems. Thus memories of what’s been seen, heard, and the like count as within the mind, but what’s written on pieces of paper, no matter how constantly close to hand, do not. Leonard’s problem is precisely that the information is not inside the mind, where it needs to be to do him any

good. The same consideration would seem to apply to Otto as well.[10]

Interestingly, Clark and Chalmers do consider the objection that what makes it the case that Otto's notes do not count as part of what he believes - as information in his mind - is the fact that he has to perceive it before he can reason with it. They do not think that this fact grounds a principled distinction between what's inside and outside the mind, because, they argue, one can just consider perception as a kind of internal channel of information flow within the mind. However, if we take seriously the differences discussed above between the global central processes that determine what we believe and the modular input processes that transform information out in the world into a form usable by central processes, I think the basis for a principled inside-outside the mind distinction exists. Leonard can't function precisely because, no matter how fast he can write, what he writes isn't in a form, or at a location, that his central reasoning can use in the particular way necessary for developing a reasonable common sense picture of the world around him. And that's why Leonard's system doesn't work.

[10]Of course there is this significant difference between Leonard and Otto. Otto's memory deficit is very small in comparison with Leonard's, so his reliance on his notebook does not put him in the impossible epistemic position that Leonard occupies. But with respect to the question whether to consider what's in the notebook as "in" his mind, the two cases are identical.

References

Antony, L.M. (1995). "Sisters, Please, I'd Rather Do It Myself: A Defense of Individualism in Epistemology." Special issue of Philosophical Topics: "Feminist Perspectives on Language, Knowledge, and Reality," ed. Sally Haslanger, Vol. 23, No. 2, 59-94.

Antony, L.M. (2003). "Rabbit-pots and Supernovas: The Relevance of Psychological Evidence to Linguistic Theory," in The Epistemology of Language, edited by Alex Barber, Oxford University Press, 2003, 47-68.

Burge, T. (1995). "Content Preservation", Philosophical Issues, Vol. 6, Content, 271-300.

Clark, A. and Chalmers, D. (1998). "The Extended Mind", Analysis 58, 10-23.

Fodor, J.A. (1983). The Modularity of Mind. Cambridge, MA: Bradford Books/The MIT Press.

Fodor, J.A. (2000). The Mind Doesn't Work That Way. Cambridge, MA: MIT Press

Ford, K. and Pylyshyn. Z. eds. (1996). The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence. Ablex Publishing.

Quine, W.V. and Ullian, J.S. (1978). The Web of Belief. McGraw Hill.

Putnam, H. 1975. "The Meaning of Meaning". In Language, Mind, and Knowledge, ed. K. Gunderson, 131-193. Minneapolis: University of Minnesota Press.

Reisberg, D. (2006). Cognition: Exploring the Science of the Mind. New York: W.W. Norton & Company

Segal, G. (2000). A Slim Book About Narrow Content. Cambridge: MIT Press.