The Compatibility of Complex Systems and Reduction: A Case Analysis of Memory Research

WILLIAM BECHTEL
Department of Philosophy, Washington University, Campus Box 1073, One Brookings Drive, St. Louis, MO 63130-4899, USA; E-mail: bill@twinearth.wastl.edu

Abstract. Some theorists who emphasize the complexity of biological and cognitive systems and who advocate the employment of the tools of dynamical systems theory in explaining them construe complexity and reduction as exclusive alternatives. This paper argues that reduction, an approach to explanation that decomposes complex activities and localizes the components within the complex system, is not only compatible with an emphasis on complexity, but provides the foundation for dynamical analysis. Explanation via decomposition and localization is nonetheless extremely challenging, and an analysis of recent cognitive neuroscience research on memory is used to illustrate what is involved. Memory researchers split between advocating memory systems and advocating memory processes, and I argue that it is the latter approach that provides the critical sort of decomposition and localization for explaining memory. The challenges of linking distinguishable functions with brain processes is illustrated by two examples: competing hypotheses about the contribution of the hippocampus and competing attempts to link areas in frontal cortex with memory processing.

No one doubts that the brain processes underlying mental activity are complex. Increasingly, researchers seeking to explain the complexity of biological systems have a range of sophisticated tools available to them. Component processes in complex systems often behave in a non-linear fashion, and the interaction of non-linear processes often generates surprising emergent phenomena. Dynamical systems theory (DST) has been put forward by some theorists as providing an innovative framework in which to understand these complex systems (van Gelder, 1995, 1998). The tools of DST can be very helpful in understanding the cognitive activity of brains just as they have proven helpful in understanding artificial neural networks that have been designed to simulate cognitive performance (Elman, 1991; Beer, 2000). A particularly interesting example is Walter Freeman's use of dynamical ideas to account for how animals detect odors (Skarda and Freeman, 1987). But there is a feature of the rhetoric of some advocates of DST that is troublesome – the repudiation of attempts to understand complex systems by analyzing or decomposing their operation into component processes and attempting to identify these component processes with physical parts of the system (what I call *mechanistic explanation* – see Bechtel and Richardson, 1993). Van Gelder (1995), for example, identifies homuncularity, the idea that one can analyze systems into components, as allied with such notions as representation, computation, and sequential and cyclic operation, all of which he views as incompatible with and supplanted by a dynamical approach. Efforts to decompose and localize processes are often ridiculed

Minds and Machines **11:** 483–502, 2001.

as reductionistic and conceived of as unable to explain the operation of complex systems.

My strategy in this paper is to argue for the compatibility of traditional reductionistic, mechanistic science with a more holistic focus on complex systems and the use of DST tools to characterize such systems. I shall proceed in two ways. I will start conceptually, focusing on the misleading manner in which the opposition to reductionism is presented. I will then show that a common outcome of pursuit of mechanistic explanation is the discovery of complex interactions between the components. These interactions then benefit from invoking the tools of dynamical systems theory. Second, I will develop a case analysis of contemporary mechanistic investigation of human memory. Current research is steadily identifying more components and interactions between components – precisely the sort of thing which provides a basis for a fruitful invocation of the tools of DST.

1. DST's Mistaken Opposition to Mechanisms

The DST movement in cognitive science (Port and van Gelder, 1995; van Gelder, 1995, 1998) approaches cognition from the perspective of the behavior of an existing behaving system, identifies variables in terms of which the system's change over time can be described, and attempts to devise differential equations involving the variables and parameters to describe these changes. Given the non-linearity of the resulting system of equations, theorists generally find it fruitful to represent the behavior of the system geometrically. To do this, one conceives of a multidimensional space, with one dimension for each variable used to describe the system. This defines the state space for the system. One can then represent the changes in the behavior of the system in terms of its trajectory through the state space. Sometimes the system will follow a trajectory to a particular point (a point attractor), sometimes it will follow a trajectory to a limit cycle (a cyclic attractor) around which it will then move. There may be more than one attractor within the state space so that from different starting points the system will settle in different location. In addition, as the parameters change, the location of attractors may change, causing the system to follow different trajectories (Betchel and Abrahamsen, in press; Kelso, 1995).

For DST theorists, the resulting accounts constitute explanations. The equations relating variables and parameters provide the explanans (as they do in covering law models of explanation). The introduction of geometric representations and of notions like attractor provide understanding of the system's behavior. This sort of explanation is contrasted with homuncular explanation which, as noted above, involves decomposing the operation of a system into subtasks and identifying each of these subtasks with different physical parts of the system.

The success of homuncular explanation, quite naturally, depends on there being humunculi within the system. Wheeler (in press; see also Clark, 1997) construes homuncular systems as a species of modular systems, which are in turn identified

with the kinds of systems Wimsatt describes as aggregative. Wheeler's construal of aggregativity is a bit different from Wimsatt's. For Wheeler

> An aggregative system is a system in which (a) it is possible to identify the explanatory role played by any particular part of that system, *without taking account of any of the other parts*, and (b) interesting system-level behaviour is explicable in terms of the properties of a small number of parts (emphasis added).

Following Clark, Wheeler construes *continuous reciprocal causation* involving "multiple simultaneous interactions and complex dynamic feedback loops" as making systems less aggregative, and then contends:

> It seems plausible that, as a system becomes less and less aggregative, with increasing continuous reciprocal causation, the more useful explanatory stance that one can take towards that system will become increasingly holistic, i.e., the most useful explanations will become increasingly *non-modular*.

Note that this is a graded claim – there is a continuum between aggregative systems and ones in which modularity fails totally. One factor leading to the gradation is that few if any systems of differentiated parts are such that the behavior of the parts is not, to some degree, influenced by the behavior of other parts. All a mechanist requires is that we can get a first approximation account of what the parts contribute by examining them individually, and then take into account the interactions. The question arises as to where neural systems involved in cognition (as well as other biological systems) are located on this continuum. DST theorists locate them near the holistic extreme. But if they are even a bit further removed from the extreme, then there may still be modules in a sufficient sense for explanation that works by assigning different functions to different components.

For a moment, let's consider how initially autonomous components might become constituents of larger systems (in biology, this may be relatively rare, but the inclusion of the mitochondrion into the cell is often suggested to be such a case). Before components will be construed as comprising a system, they are required to participate in appropriate interactions. Unless we have designed a system by physically locating components in a common setting, there is little reason to construe simple linear interactions, such as when component A sends its output to unit B, as making the components part of a system. But when the operation of A is itself partially dependent upon that of B as well as B utilizing the output of A (thus violating the italicized clause in the quotation from Wheeler above), the idea that A and B comprise one system becomes more compelling. Interestingly, DST theorists, despite their antihomuncular rhetoric, have a term for such situations: *coupled dynamical systems*. The degree of coupling further determines the usefulness of construing A and B as parts of a common system or of continuing to focus on the systems that are coupled.

I have focused here on functional coupling between components. But frequently functional coupling is linked to physical coupling – if components depend on each

other, then it is useful for them to be located in close physical proximity, and there will be evolutionary pressure to keep them together. Thus, it is not surprising to find in cells that the different enzymes needed for respiration enclosed inside the mitochondrion, those needed for protein synthesis in the ribosome, and those used for breaking down old cell components in the lysosome. An important feature of being incorporated into a complex system is that the components loose part of their autonomy and their behavior becomes modulated and regulated by the environment (other components) within the system. Claude Bernard introduced the idea of an internal milieu to describe the interior of an organism and emphasized the importance for life of maintaining the constancy of the internal milieu. The result is that components behave appropriately for the conditions within that environment, not as they might outside that environment.

But now we can consider the critical point – does the fact that components become coupled into systems as a result of increased interactions and dependencies between them, and the fact that their behavior is partly regulated by these interactions, entail that we can no longer usefully apply homuncular analysis to them? At the level at which such coupling appears in biological systems, the answer seems to be "no." Even if mitochondria were once independent structures that become incorporated into cells and became so linked to other parts of cells that they cannot survive in isolation, that does not mean that we cannot analyze what distinctive contributions they make to cells as well as how their operation is modulated by other cell components. The quest for mechanistic explanation through decomposition and localization that Richardson and I described involves discovery not just of localized components but of considerable interactions between them. What we characterized as *integrated systems* involved feedback loops which linked the behavior of individual components to that of others.

Above I noted that biological systems are typically not composed from previously autonomously systems being incorporated into a common system. The more usual pattern involves differentiation within an existing system. What was once one part is divided into two parts, which then differentiate further over time. This looks to be the way in which distinguishable cortical areas have developed in the brain. Even though the areas never functioned independently, the existence of differentiated cortical areas is a powerful argument for homuncularism. Once microscopes and staining techniques were sufficiently developed, researchers in the first decades of the 20th century began analyzing the neural composition of the cerebral cortex and noting differences between cortical areas. Sometimes the differences turned on the proportion of different types of neurons; other times they turned on the distribution of neurons. On staining, six major layers of cells can be differentiated across cortex, but the thickness and subdivisions within these layers varies in different parts of cortex. Brodmann's famous map showing 47 different brain areas in the human (several other researchers in the same period advanced different maps) was based on such physical differences, although Brodmann makes it clear that his objective was to distinguish areas that were performing different functions.

Although Brodmann's designations are still widely used, research on cortical mapping today usually makes far finer differentiations. Felleman and van Essen (1991), for example, distinguish 33 different areas involved in visual processing alone in the macaque. A variety of different techniques figure in establishing these areas, but ultimately the interest is in determining what each area contributes to mental activity. Accordingly, van Essen and Gallant (1994) link a variety of subtasks of visual processing with different areas. I take their account to be an exemplar of homuncular mechanistic analysis of how the brain performs a cognitive function (Bechtel, 2001).

One of the keys both to differentiating areas and to developing conceptions of how brain areas interact in mental function is studying the connectivity between cells in different areas. Although the number of connections is large, it is relatively orderly – only about a third of possible area to area connections between the 33 visual areas identified by Felleman and van Essen are actually realized, permitting the development of a complex mapping of the relations between areas. Moreover, in several cases the mappings between areas preserve the topographical representations in each area. In most cases where there are forward connections, there are also reciprocal connections. Forward, backwards, and lateral connections are typically distinguishable in terms of the layers from which they originate and layers in which they terminate. Overall, the visual system seems to be a complex system, but one in which there are distinct areas carrying out different operations on visual input which interact with each other in a structured manner.

The fact that the visual system and other brain systems are homuncular and can be understood in mechanistic terms, however, does not mean they are not also fruitfully characterized in dynamical terms. As just noted, the brain areas involved in vision are highly interconnected. The behavior of different components is constrained by their relations to other components. Given the vast number of connections, feedforward, feedback, and collateral, tools such as DST affords are likely to be extremely useful. (The use of such tools in the analysis of much simpler artificial neural networks provides a useful model of how they might be applied. See Hinton and Shallice, 1991; Plaut et al., 1994). The point to be emphasized, though, is that the tools are then employed in the context of a homuncular analysis, and are not in conflict with it.

2. Decomposing and Localizing Memory

Having argued that dynamical tools are compatible with decomposition and localization and that their most important use may be in understanding humuncular system with complex interactions, I turn now to an example of how the program of decomposition and localization plays out. I have already pointed to vision as a cognitive ability where research has identified a host of parts and interconnections and where the stage is set for fruitful utilization of dynamical tools. But the challenges in developing decompositions and localizations are often not appreciated and fre-

quently it is the shortcomings of early stages in developing such explanations that engender the opposition. To exhibit some of the different facets of such research and how difficult it is to develop the homuncular analysis that is foundational to employing the tools of DST, I will focus on contemporary research on memory.

A common first move of theorists trying to understand a phenomenon is to attribute the phenomenon to a component in the larger system in which it appears. Since memories are clearly something animals, especially humans, retain, a natural question to ask is where they are stored: *where is the engram*? The difficulty of finding brain locations that encoded specific memories convinced Lashley among others that the search for an engram was misguided. Lashley ended up endorsing a holism in many ways reminiscent of the holism advocated by contemporary proponents of DST. But engrams and holism are not the only alternatives. The challenge in opening up other alternatives is to find fruitful ways of decomposing the system.

Working almost exclusively with behavioral tools, cognitive psychologists have advanced a number of decompositions of memory phenomena based on such things as how much can be remembered and the sensitivity of memories to interference. Differences in capacity in part underlay the distinction between short-term and long-term memory advanced in the 1950s and 1960s (Waugh and Norman, 1965). Miller (1956), for example, presented evidence that the capacity of short-term memory (what one can remember for seconds or minutes as long as one is not interrupted) was 7 plus or minus three items. In contrast, the capacity of long-term memory is essentially unlimited. For extremely short periods, the capacity limitation of short-term memory also does not apply. If one is briefly presented with a table of several rows of numbers and, immediately after the display is removed, asked to recite a specific row, one can generally do so, but then loses access to all the others (Sperling, 1960). This extremely short-term retention was characterized as echoic memory. Not only were three different sorts of memory distinguished, but in an early information processing model, Atkinson and Shiffrin (1968) proposed that echoic memory supplied inputs to short-term memory, which in turn supplied inputs into long-term memory, with attention modulating whether information was passed from one type of memory to the next.

The Atkinson and Shiffrin model is a characteristically homuncular analysis of the sort that has frequently been advanced in cognitive psychology. But the study of memory soon became much more complex. Already in the 1950s memory became one of the cognitive capacities in which both psychological and neuroscientific perspectives were brought together as a result of surgery William Scoville performed on a patient that has come to be known as HM. To relieve HM's severe epileptic seizures, in 1953 Scoville removed much of his medial temporal lobe. The surgery was successful in relieving HM's epilepsy, but left him severely amnesic. In particular, he has acquired no new memories of episodes in his life since the surgery (anterograde amnesia), and has graded loss of memory for events for several years preceding the surgery (retrograde amnesia) (Scoville and Milner, 1957). To account

for these results with HM, researchers advanced the hypothesis that the hippocampus plays a critical role in memory. But equally important was the discovery that HM could acquire new skills even though he had no memory of learning them (Corkin, 1968). Subsequently, numerous other patients have been identified whose amnesias are interestingly different from HM's. KC, studied by Endel Tulving, retains no memory for any events in his life subsequent to a motorcycle accident in the early 1980s, but can learn, albeit with difficulty, new factual information, which HM cannot (Tulving et al., 1991). These discoveries, as well as advances in purely behavioral research on normal individuals, has generated numerous taxonomies of types of memory. It has also resulted in a controversy over whether these types of memory ought to be attributed to different memory systems. As I shall argue, this decomposition into memory systems, while it seems to be the kind of decomposition that supports mechanistic explanation insofar as it is engaged in decomposing and localizing memory systems, is actually orthogonal to the development of mechanistic explanation. It serves rather to differentiate phenomena that require explanation. Accordingly, I will refer to the decomposition involved as *phenomenal decomposition* and distinguish it from *mechanistic decomposition*.

A. Memory systems versus memory processes

In 1972 Tulving introduced a distinction between "two parallel and partially overlapping information processing systems" (p. 401), one (*episodic memory*) concerned with episodes in our lives, which specifies information about the time and place of their occurrence, the other (*semantic memory*) concerned with memory of general information (word meanings, scientific facts) which is generally retrieved independently of recalling the time and place in which it was acquired. Both are species of long-term memory. Tulving proposed that different types of tests measured episodic and semantic memory performance – recall and recognition of recently studied events for episodic memory versus retrieval of a word from a fragment, retrieval of a word from its definition, identifying words from brief tachistoscopic displays, and lexical decisions (i.e., deciding whether a letter string constituted an English word) for semantic memory.

In 1980 Cohen and Squire (1980) advanced a different decomposition of memory that corresponded to Ryle's (1949) distinction between types of *knowledge – knowing how* and *knowing that*. Knowing how involves mastery of activities like riding a bicycle or constructing logic proofs, and Cohen and Squire termed the retention of this information *procedural memory*. Knowing that, on the other hand, involves knowledge that could be explicitly stated, and Cohen and Squire termed this *declarative memory*. Tulving's distinction between episodic and semantic memory was then construed as marking subtypes of declarative memory.

The differentiation of different types of long-term memory raises an important question as to what the differentiation ultimately comes to. Initially, Tulving claimed that his distinction between episodic and semantic memory was "primarily for the convenience of communication, rather than as an expression of any

profound belief about functional structure separation of the two" (1972, p. 384). Shortly thereafter, however, Tulving became an advocate for the claim that these represented different *memory systems*. Although there are a number of advocates, especially among those strongly influenced by evidence from lesion cases that exhibit dissociation of different types of memory, of the multiple system approach, there is no consensus about what constitutes a *system*. One common feature of many conceptions of a system is that they (a) are, in some sense, structurally distinct, (b) process different types of information and represent it differently, and (c) operate in accord with different principles (for an early statement, see (Tulving, 1984). In a recent statement of the position, Tulving distinguishes memory systems partly in terms of the type of information represented, using as an example a person who has read the sentence "aardvarks eat ants":

> PRS, the perceptual representation system [an additional system Tulving introduced more recently], encodes and stores information about the features of the visual objects represented by the letter strings AARDVARKS EAT ANTS. The semantic memory system, or a set of its (presumably numerous) subsystems, encodes and stores propositional information about the feeding habits of animals named aardvarks. The episodic system integrates, registers, temporally dates, and spatially localizes the rememberer's experience of the experience of being present and witnessing the sentence appearing on and disappearing from the screen (p. 20).

Schacter and Tulving (1994) argue for three criteria for a memory *system*: (1) a system "enables one to perform a very large number of tasks of a particular class or category, regardless of the specific informational content of the tasks" (p. 15), (2) a system must be described by a set of properties (e.g., rules of operation, kind of information handled, and neural substrates) as well as in terms of "what the system is for" (p. 16), (3) a system is distinguished form others by "converging dissociations: dissociations of different kinds, observed with different tasks, in different populations, and using different techniques" (p. 18).

The strategy of researchers adopting the systems approach is clear. Memory is decomposed into systems on the basis of differences in the way different kinds of memory function, the properties they exhibit, etc. These different systems are connected to the brain by discovering brain areas that, when lesioned, destroy the kind of memory in question or that can be shown to be particularly active when people exhibit that kind of memory. Things become more complex since some brain areas may figure in different kinds of memory and thus be assigned to multiple systems. But what I want to emphasize here is that the decomposition is in terms of the phenomena to be explained – different types of memory exhibit different properties and require different explanations.

The claims for multiple memory systems have not gone uncontested in the psychological literature, although there is substantial disagreement over what the opposition is disputing. Tulving construes the debate as one over multiple versus a single memory system, and construes advocates of a single memory system as

holding that a common mode of representation is employed for all memory tasks. But the opponents of memory systems have often adopted a different perspective, arguing for memory processes as opposed to memory systems. Roediger, Buckner and McDermott characterize the process approach as follows:

> The hallmark of the procedural approach, harking back to Bartlett and Neisser, was that performance on memory tasks could be described as skilled performance and that one should look to the procedures of mind to explain cognitive performances. Many experiments can be interpreted as supporting the procedural approach, including several revealing dissociations in performance on tasks that all measured recognition of words. In particular, Kolers' experiments showed that transfer from one task to another benefitted to the degree that the procedures underlying performance on the two tasks were similar (Roediger III et al., 1999, p. 42).

Although on first appearances it may seem as if the difference between multiple processes and multiple systems is only terminological, it is in fact a fundamental difference. This appearance appears minimized by the fact that dissociation, the most potent tool for distinguishing multiple memory systems, in often used to distinguish different processes. For example, in their work on transfer appropriate processing, Bransford et al. (1979) showed that if an encoding task emphasized either surface or meaningful features of the item to be encoded, recall would be better when the recall task emphasized the same features. But there is a crucial difference: multiple processes are construed as multiple steps in a stream of processing steps, not as comprising independent systems. Multiple systems operate independently of each other (they are similar to Fodorian modules) whereas multiple processes interact and combine to perform cognitive operations. The dissociations found in transfer appropriate processing studies are all within what proponents of the systems view construe as the episodic memory system, leading Kolers and Roediger (1984) to comment: "If dissociations are found among tests tapping the same memory system, then the discovery of dissociations between tasks cannot be taken as evidence for different memory systems" (p. 438).

When early processing theorists offered a distinction of different types of processing, they began with a distinction between bottom-up (or perceptually based) processing and top-down (or conceptually driven) processing. Since most tests of implicit memory can also be construed as tests of bottom-up processing, and most tests of explicit memory as tests of top-down processing, it is difficult to determine experimentally whether this proposed difference in types of processing is empirically better supported than the differentiation of systems. When researchers have surmounted this difficulty (e.g., by developing tests that involve perceptual processing and explicit recall), the resulting data was ambiguous. Blaxton (1989), for example, produced data in which generating words from conceptual cues as opposed to simply reading the words led to better recall on conceptual tests than perceptual tests, whether episodic or semantic, whereas reading produced better recall on perceptual tests than conceptual ones, whether episodic or semantic. While

these seem to favor the processing account, Tulving and Schacter (1990) used them to revise the systems account (adding the perceptual representation system noted above). On the other hand, McDermott and Roediger (1996), themselves advocates of the processing approach, produced evidence dissociating different conceptual tasks, indicating that top-down processing does not always rely on the same processes.

According to Roediger et al. (1999), both the traditional process theories and the traditional systems theories fail to accommodate all the data. But they contend that an alternative framework, *the components of processing framework*, developed by Morris Moscovitch, provides a more adequate framework that can resolve the conflict between the approaches. As the term *processing* in its name suggests, the components of processing framework is a descendent of the memory processes approach, with the additional idea that different tasks may draw differentially upon different components in a processing system. If two tasks can be dissociated (a manipulation can affect performance on one task but not on the other), then there must be at least one component process that figures differently in the two tasks (Hintzman, 1990). Within this framework, dissociations are no longer used to tease apart whole systems, but only differences in reliance on components within a larger system. It is here that the distinction between the systems and process approach becomes sharp. It also becomes clear that the process approach embraces mechanistic decomposition, for in this approach researchers are engaged in identifying the different information processing activities that are recruited in the performance of a memory task.

Roediger et al. use neuroimaging for three different variations of a word stem completion task to illustrate the approach. In the baseline condition, subjects were just asked to complete a stem like COU with the first word that came to mind (a purely semantic memory task). In the other two conditions, the subjects first study a list of words. In the second condition, subjects are then given the same directions as on the word stem completion task – complete the stem with the first word that comes to mind, while in the third condition, they were instructed to only complete the stem with a word they had just studied (an explicitly episodic memory task). In the first condition, word-stem completion with no previous exposure to words, increased activation was found in areas of visual cortex bilaterally, left frontal opercular cortex and supplementary motor areas, right premotor cortex and anterior cingulate. When the words were primed by prior exposure, the same areas were activated, but with reduced activations in visual areas, which the researchers interpret as evidence of incidental memory retrieval. When instructions were added to complete the stems only with previously primed words, all the areas activated in the previous conditions were again activated, as well as two additional areas: anterior prefrontal cortex bilaterally and posterior medial parietal cortex. These are areas that have also been activated in other studies of episodic memory, but rather than construing this as evidence for a separate system for episodic memory, Roediger et al. interpret it as evidence for multiple components of a broader memory

system that are employed when the subject is asked to evaluate whether the items were previously encountered.

Since the systems approach does allow separate systems to share components, its advocates could accommodate Roediger et al.'s results. The systems approach can simply define as a system whatever components figure in, for example, episodic memory tasks. The challenge for the system approach gets more difficult when different neural components are involved in different tasks that all seem to involve episodic memory, since there then seems to be little integrity to the proposed episodic memory system. But rather than focusing on the question of empirical adequacy, I will focus on the underlying understanding of the task of scientific explanation advanced by the two approaches. The systems approach separates systems according to the phenomena to be explained – episodic memory, semantic memory, etc. This is an important differentiation, but its primary merit is not in developing explanations of *how* memory works – rather, it is a preliminary task of determining what kinds of memory activities need to be explained. The process approach and its descendent, the components of processing view, involves decomposing memory in a very different manner than the memory systems approach. The process approach differentiates components within the system in terms of what kind of information processing each performs. The process approach is truly engaged in mechanistic explanation.

B. Assigning memory functions to brain regions

Starting with the investigation of HM, neurobiologically oriented memory researchers have identified a number of brain regions that seem to be involved in one or another type of memory. Working within the processing framework discussed in the last section, the challenge is typically not to show that these areas play a role in memory, but to determine what information processing activity they perform. In this final section I will examine how the attempt to link processes with brain structures has played out with respect to two different brain areas, the hippocampus and surrounding medial temporal lobe, which was damaged in HM, and prefrontal areas which Tulving proposed were involved in encoding and retrieval of episodic memories.

The hippocampus has been a popular focus of theorizing because of its distinctive architecture. It consists of several different areas with very different composition. Input to the hippocampus is funneled through the parahippocamal regions of the temporal lobe and then through the entorhinal cortex (EC) into the hippocampus itself. The hippocampus consists of a loop involving the dentate gyrus (DG), CA3 region, and CA1 region (see Figure 1). Inputs from the EC project both to the DG and CA3 along the perforant pathway. The DG consists of granule cells, only a few of which fire at a time Each granule cell sends projections along the mossy fibers to just a few CA3 cells. CA3 is comprised of pyramidal cells which are highly interconnected via recurrent connections. CA3 cells send projections to the more

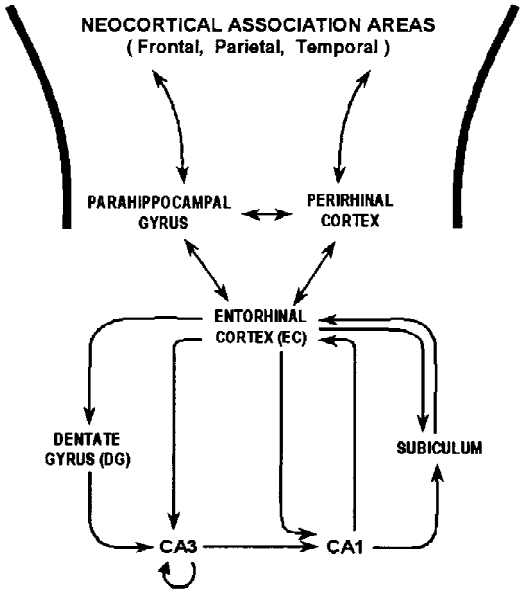

Figure 1. A schematic representation of the major input pathways into the hippocampus (top) and the connectivity within the hippocampus itself (bottom).

numerous pyramidal cells in CA1 via the Schaffer collaterals (CA1 also receives input directly from EC. CA3 cells project back to the EC either directly or via the subiculum, completing a functional loop. The distinctiveness of this neuroanatomy suggests that the hippocampus subserves a cognitive function that could not be performed using more typical cortical architecture. What is interesting is the radical disagreement among researchers as to what this function is.

The case of HM strongly suggested that the hippocampus has a role in acquiring declarative memories. It also shows that the hippocampus is not the site of permanent memory storage, since HM retains memories from several years prior to his surgery. One of the most intriguing features, though, is his retrograde amnesia for the years immediately prior to his surgery. Such retrograde amnesia has been found in numerous other patients, leading researchers to infer that new episodic memories are being temporarily encoded in the hippocampus, but then transferred to other parts of cortex. Numerous theorists (McClelland et al., 1995; Rolls and Treves, 1998) have proposed that the hippocampus both provides for temporary storage of new memories and serves as the trainer for other cortical regions, which gradually acquire the memory as the hippocampus repeatedly reinstates the neural pattern that encodes the memory. (The corollary argument is that the neocortex could not acquire new memories directly since the rapid learning rate this would require would produce catastrophic interference – the rapid loss of previous memories as new memories are acquired.) This construal of the hippocampal function suggests a role for the overall loop through the hippocampus – it could provide an autoasso-

ciator that, on the basis of similar experiences could reinstate patterns produced by previous experiences, a process that has obvious relevance for a memory system. A similar significance is assigned to the recurrent loops within CA3 itself. But there is a disadvantage to such an autoassociator – because it is good at reinstantiating the same response to similar inputs, it is incapable of separating distinct experiences. The sparse connectivity pattern in the DG, on the other hand, suggests that it could play a role in pattern separation. Both McClelland et al. and Rolls and Treves have developed models showing how the hippocampus could play a critical role in temporary encoding of declarative memories and, through the ability to reinstantiate these memories, train other areas of cortex, which then become their permanent repository.

In the same time period as when research on amnesia in humans was supporting a role for the hippocampus in encoding episodic memories, research on lower mammals, especially mice and rats, suggested a very different function, a role in spatial memory. From studying the maze running behavior of mice, Tolman had proposed that they developed cognitive maps of their spatial environment that enabled them to solve spatial navigation tasks. O'Keefe and Nadel (1978) discovered that lesioning the hippocampus in rats impaired some of these navigational abilities. In the Morris water maze, for example, the target is a submerged platform on which the rat can stand and a normal rat released from a different location from where it first found the platform will swim directly towards it. Without a hippocampus, the rat searches anew for the platform each time (Morris et al., 1982). It does not reveal the deficit if it is regularly released from the same starting point, presumably because it can rely on a memorized route. Without a hippocampus, the rat can still navigate by landmarks (what O'Keefe and Nadel term *taxon* navigation). To swim directly to the submerged platform from novel locations, they propose, the rat requires an allocentric representation of space (that is, a representation not based on relations to itself) and then must be able to represent its current location in that allocentric representation and plot an appropriate route. That the hippocampus provides such an allocentric representation is further supported by single-cell recording studies which have identified cells in CA3 that fire when the rat is in particular location in an environment (these cells are called *place cells*), suggesting that they carry information about where the rat is in allocentric space (O'Keefe and Dostrovsky, 1971). In recent years numerous computational modelers have tried to account for the capacities of place-cells computationally (Zipser, 1985; Hetherington and Shapiro, 1993; Touretzky and Redish, 1996). In Touretzky and Redish's model, that loops in CA3 provide attractors that facilitate reactivation of the same cells when a rat returns to a previous location, whereas the sparse pathway from DG to CA3 serves to distinguish different locations.

The assignment of different functions to the hippocampus has led to active competition between the proponents of the two approaches, each of which retains devoted advocates (Nadel, 1994). What is interesting is that when advocates of each approach advance proposals as to how the hippocampus can perform the preferred

function, they appeal to the same features of the hippocampus's neuroarchitecture. If the same neuroarchitecture would indeed be useful for both functions, that suggests that the two functions might, after all, be compatible, and both performed by the hippocampus. Accordingly, some researchers have tried to bridge between the two approaches. From the animal navigation side, Redish suggests how a role in spatial encoding provides the foundation for the hippocampus to play a role in encoding declarative memory. Starting from the focus on declarative memory, Eichenbaum (Cohen and Eichenbaum, 1993; Eichenbaum et al., 1993) proposes that what is crucial about declarative memory (and what the hippocampus accomplishes) is establishing relationships between information items that can be accessed in a flexible manner. He proposes that spatial memory is just one example of such a kind of memory.

As a brain region involved in memory, the hippocampus is interesting in that two seemingly very different conceptions of its function emerged from two different research traditions, and subsequent research has focused on either vindicating one of the conceptions or trying to advance an account that reconciles them. But another, and very common pattern in developing accounts of complex systems, is to start with an assignment of a function to a brain region, and then modify it as further research is done. This pattern is exemplified in research on prefrontal cortex. Initially prefrontal cortex did not seem to play a central role in memory since patients with prefrontal lesions did not seem to exhibit memory deficits. One of the first studies to direct attention to this brain area was a PET study by Tulving and his colleagues (Tulving et al., 1994). Tulving was led to apply PET to study memory since it, unlike purely behavior or lesion studies, offered the potential to separate encoding and retrieval processes. Any behavioral measure of memory requires that the subject both encode and retrieve the items to be remembered, whereas in neuroimaging one can look separately for brain regions exhibiting increased activation during encoding and retrieval. To measure involvement in encoding, Tulving et al. compared encoding tasks that encouraged shallow processing with tasks that required deep processing (which results in better performance on recall tasks). The result was increased activation in left dorsolateral prefrontal cortex with deep processing. Imaging during recall, on the other hand, resulted in increased activation in right anterior prefrontal cortex. From these results Tulving advanced the hemispheric encoding/retrieval asymmetry hypothesis (HERA).

HERA represents a bold hypothesis about brain organization, one that encourages further experimentation that will either contest it or support it. Much of the evidence points to a more complex distribution of tasks. Buckner (1996), for example, found that processing areas in dorsolateral left prefrontal cortex that are active in retrieval in semantic memory tasks (stem completion) are also active in episodic memory retrieval (requiring stem completion from words on a previously studied list). There is a good explanation why this activation was not noted in the initial development of HERA – it was concealed by the subtraction technique used in Tulving's and many other early neuroimaging study in which activations

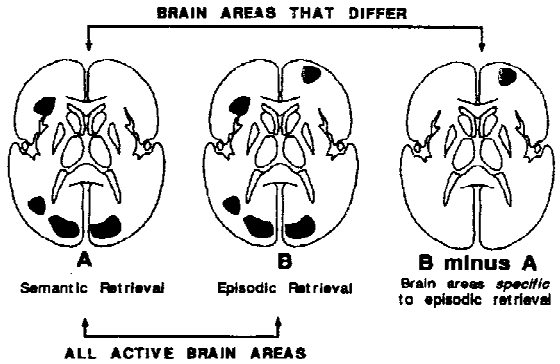

Figure 2. Buckner's (1996) representation of brain areas active in (A) a semantic retrieval task and (B) an episodic retrieval task. In both of these, what is subtracted is activations in a low level control task which makes minimal memory demands. When the activations in the semantic retrieval task are subtracted from those in the episodic retrieval task (C), only the area differentially involved in episodic retrieval is activated. But the episodic retrieval task required all the areas shown in (B), not just the right prefrontal area featured in (C).

produced by one task were subtracted from those produced in other task. This is useful if the goal is to discover what additional brain areas might be involved in a given task; but if the goal is to develop a mechanistic model decomposing tasks into component processes and detailing their interactions, then one needs to identify the whole set of areas actually involved in the task. Buckner's analysis showed both left dorsolateral prefrontal cortex and right anterior prefrontal cortex are involved in episodic retrieval (see Figure 2), although it was not able to determine what information processing was carried out by these or the other active areas.

A new approach to neuroimaging, event related fMRI, has recently provided a means to identify the differences in brain activations during encoding of stimuli the subject later remembers and those the subject fails to remember later. Utilizing this technique with an encoding task in which subjects were required to make a concrete versus abstract judgment, Wagner et al. (1998) found activation in several left prefrontal areas (posterior and anterior left inferior prefrontal gyrus and left frontal operculum) on successful encoding trials. This basically accords with HERA, but a similar study by Gabrieli and his colleagues demonstrated right inferior frontal cortex activation, as well as bilateral hippocampal activations, when the stimuli were pictures instead of words (Brewer et al., 1998). Right hemisphere activation on encoding conflicts with HERA and suggests that the left-right asymmetry might be more related to whether the stimuli are words versus pictures than encoding versus retrieval.

Kelley et al. (1998) likewise produced evidence of right hemispheric activity in encoding of unfamiliar faces, and bilateral activations with line-drawings of nameable objects, indicating that the left hemispheric activations might be more the product of activating of semantic processing areas as the subject supplies linguistic labels. Such a result seems to seriously challenge HERA, but see Nyberg

et al. (1998) for a response. The fact that the dorsolateral left prefrontal areas that are activated in these studies are close to areas in which increased activation was observed in studies of semantic processing of words (Petersen et al., 1989), though, suggests a potentially even more substantial challenge to these forays into decomposing and localizing memory processes. As Gabrieli et al. (1998) propose in a related context, perhaps the initial phenomenal decomposition of memory and language processing is mistaken. Similar information processing operations might figure in both, and the brain might not decompose mental activity in the same manner as human researchers.

The examination of prefrontal areas in memory is at a very early stage. This brief look at conflicting ideas about what sort of processing these prefrontal brain areas perform as well as the previous consideration of conflicting ideas about what the hippocampus might contribute to memory, though, both shows how researchers engaged in the quest for understanding the neural mechanisms of cognition proceed and the challenges confronting them. All these researchers are strongly committed to decomposition and localization of function. But there are two procedures for decomposing memory – decomposing into different kinds or systems of memory and decomposing into different cognitive operations involved in memory. Both give rise to localization, but in the former case, which I have referred to as phenomenal decomposition, the localized systems are often overlapping and there is a serious danger of failing to identify the various interactions between the localized areas. The second strategy, process or mechanistic decomposition, opens the potential for discovering complex, interactive and integrated systems. Often the tools for localizing – relying on deficits from lesions or subtractions between neuroimages – can mislead one to think more is localized in a given location than actually is. But this danger is often temporary, and continued pursuit can lead to discovering multiple components with complex patterns of interaction. This is the stage research is entering, both in the case of the hippocampus and in the case of frontal regions.

Advocates of dynamical systems theory might portray the difficulties memory researchers face in determining just what each component contributes as indicating the failure of the strategy of decomposition and localization (van Orden and Paap, 1997; Van Orden et al., in preparation). In a host of other cases in science, though, initial difficulties in developing convincing accounts of what components in a system were doing have given way to well worked out analyses (Bechtel and Richardson, 1993). Moreover, as I suggested in the first part of this paper, the development of plausible hypotheses about the distinctive contributions of different parts has provided the foundation for fruitful application of dynamical models. Seeking dynamical models in the absence of a program of decomposition and localization, on the other hand, may produce vacuous science.

3. Conclusion

Many biological systems are complex, and the brain is certainly an example of a complex system. But complexity is compatible with different components performing different functions – with what fans of DST sometimes dismiss as homuncularism. The tools advocated for analyzing complex dynamical systems are indeed important, but they are most useful when combined with the results of mechanistic analysis through decomposition and localization. When successful, that approach reveals multiple components in the system (areas in the brain). Each performs a different information processing activity, but often the behavior of one area is influenced by activity in other areas. It is when they are applied to the results of such reductionistic, mechanistic research that DST's tools for analyzing complex systems are likely to be most fruitful.

Notes

[1]What is constructed as reduction in actual science often has little to do with theory reduction as it is characterized in much phiolsophical literature. I use the term *reduction* in the manner of scientists who count any attempt to understand systems by discovering their parts and the contributions those parts make.

[2]For Wimsatt, differentation of components so that they have different structures and perform different functions is already a step away from aggregativity.

[3]The term *engram* was introduced by Richard Semon, who basically presented the current view of memory as involving encoding, storage (engram), and retrieval (Schacter et al., 1878).

[4]Scoville thought he had removed HM's hippocampus, but later neuroimaging studies revealed that the lesion spared much of the hippocampus proper, although it subsequently atrophied as a result of loss of its normal inputs from surrounding cortical structures (Corkin et al., 1997).

[5]In discussion of the hippocampus there is considerable lack of clarity as to whether researches are referring to the hippocampus itself or the complex that involves several of these surrounding areas.

[6]In earlier studies, Tulving used regional cerebral blood flow to prefrontal areas involved in retrieval of episodic memories from more posterior areas involved in retrieval of semantic memories (Tulving, 1989).

[7]Recall of semantic memories also resulted in increased activation in left dorsolateral prefrontal cortex. This is explained by the fact that the tasks that produce deep encoding are precisely those that access semantic information (e.g., analysis of the meaning of the word).

References

Atkinson, R.C. and Shiffrin, R.M. (1968), 'Human memory: A proposed system and its control processes', in K.W. Spence and J.T. Spence, eds., *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 2, pp. 89–195, New York: Academic.

Bechtel, W. (2001), 'Decomposing and localizing vision: An exemplar for cognitive neuroscience', in W. Bechtel, P. Mandik, J. Mundale and R.S. Stufflebeam, eds., *Philosophy and the neurosciences: A reader*, Oxford: Basil Blackwell.

Bechtel, W. and Abrahamsen, A. (in press), *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks*, Second Edition, Oxford: Basil Blackwell.

Bechtel, W. and Richardson, R.C. (1993), *Discovering complexity: Decomposition and localization as scientific research strategies*, Princeton, NJ: Princeton University Press.

Beer, R.D. (2000), 'Dynamical approaches to cognitive science', *Trends in Cognitive Sciences* 4, pp. 91–99.

Blaxton, T.A. (1989), 'Investigating dissociations among memory measures: Support for a transfer appropriate processing framework', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15, pp. 657–668.

Bransford, J.D., Franks, J.J., Morris, C.D. and Stein, B.S. (1979), 'Some general constraints on learning and memory research', in L.S. Cermak & F.I.M. Craik, eds., *Levels of processing in human memory*, Hillsdale, NJ: Erlbaum, pp. 331–354.

Brewer, J.B., Zhao, Z., Desmond, J.E., Glover, G.H. and Gabrieli, J.D.E. (1998), 'Making memories: Brain activity that predicts how well visual experience will be remembered', *Science* 281, pp. 1185–1187.

Buckner, R.L. (1996), 'Beyond HERA: Contributions of specific prefrontal brain areas to long-term memory retrieval', *Psychonomic Bulletin and Review* 3(2), pp. 149–158.

Clark, A. (1997), *Being there*, Cambridge, MA: MIT Press.

Cohen, N.J. and Eichenbaum, H. (1993), *Memory, amnesia, and the hippocampal system*, Cambridge, MA: MIT Press.

Cohen, N.J. and Squire, L.R. (1980), 'Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that', *Science* 210, pp. 207–210.

Corkin, S. (1968), 'Acquisition of motor skill after bilateral medial temporal-bole excision', *Neuropsychologia* 6, pp, 255–265.

Corkin, S., Amaral, D., Gonzalez, R. and Johnson, K. et al. (1997), 'H.M.'s medial temporal lesion: Findings from magnetic resonance imaging', *The Journal of Neuroscience* 17, pp. 3964–3979.

Eichenbaum, H., Otto, T. and Cohen, N.J. (1993). 'Two component functions of the hippocampal memory systems', *Behavioral and Brain Sciences* 17(3), pp. 449–472.

Elman, J.L. (1991), 'Distributed representations, simple recurrent networks, and grammatical structure', *Machine Learning* 7, pp. 195–224.

Felleman, D.J. and van Essen, D.C. (1991), 'Distributed hierarchical processing in the primate cerebral cortex', *Cerebral Cortex* 1, pp. 1–47.

Gabrieli, J.D.E., Poldrack, R.A. and Desmond, J.E. (1998), 'The role of left prefrontal cortex in language and memory', *Proceedings of the National Academy of Sciences, USA* 95, pp. 906–913.

Hetherington, P.A. and Shapiro, M.L. (1993), 'A simple network model simulates hippocampal place fields: 2. Computing goal directed trajectories and memory fields', *Behavioral Neuroscience* 107, pp. 434–443.

Hinton, G.E. and Shallice, T. (1991), 'Lesioning a connectionist network: Investigations of acquired dyslexia', *Psychological Review* 98, pp. 74–95.

Hintzman, D.L. (1990), 'Human learning and memory: connections and dissociations', *Annual Review of Psychology* 41, pp. 109–139.

Kelley, W.L., Miezin, F.M., McDermott, K., Buckner, R.L., Raichle, M.E., Cohen, N.J. and Petersen, S. E. (1998), 'Hemispheric specialization in human dorsal frontal cortex and medial temporal lobes for verbal and nonverbal memory encoding', *Neuron* 20, pp. 927–936.

Kelso, J.A.S. (1995), *Dynamic patterns: The self organization of brain and behavior*, Cambridge, MA: MIT Press.

Kolers, P.A. and Roediger III, H.L. (1984), 'Procedures of Mind', *Journal of Verbal Learning and Verbal Behavior* 23, pp. 425–449.

Lashley, K.S. (1950), 'In search of an engram', *Symposium on Experimental Biology* 4, pp. 45–48.

McClelland, J. L., McNaughton, B. L. and O'Reilly, R. C. (1995). 'Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory', *Psychological Review* 102(3), pp. 419–457.

McDermott, K.B. and Roediger III, H.L. (1996), 'Exact conceptual repetition dissociate conceptual memory tests: Problems for transfer appropriate processing theories', *Canadian Journal of Experimental Psychology* 50, pp. 57–71.

Miller, G.A. (1956), 'The magical number seven, plus or minus two: some limits on our capacity for processing information', *Psychological Review* 63, pp. 81–97.

Morris, R.G.M., Garrud, P., Rawlins, J.N.P. and O'Keefe, J. (1982). 'Place navigation impaired in rats with hippocampal lesions', *Nature* 297, pp. 681–683.

Mundale, J. (1998), 'Brain mapping', in W. Bechtel and G. Graham, eds., *A companion to cognitive science*, Oxford: Basil Blackwell.

Nadel, L. (1994), 'Hippocampus, space, and relations', *Behavioral and Brain Sciences* 17, pp. 490–491.

Nyberg, L., Cabeza, R. and Tulving, E. (1998). 'Asymmetric frontal activation during episodic memory: what kind of specificity?' *Trends in Cognitive Sciences* 2, pp. 419–20.

O'Keefe, J. and Dostrovsky, J. (1971), 'The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat', *Brain Research* 34, pp. 171–175.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*, Oxford: Clarendon Press.

Petersen, S.E., Fox, P.J., Posner, M.I., Mintun, M. and Raichle, M.E. (1989), 'Positron emission tomographic studies of the processing single words', *Journal of Cognitive Neuroscience* 1, pp. 153–170.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. and Patterson, K.E. (1996), 'Understanding normal and impaired word reading: Computational principles in quasi-regular domains', *Psychological Review* 103, pp. 56–115.

Port, R. and van Gelder, T. (1995), *It's about time*, Cambridge, MA: MIT Press.

Redish, A. D. (1999). *Beyond the cognitive map: From place cells to episodic memory*, Cambridge, MA: MIT Press.

Roediger III, H.L., Buckner, R.L. and McDermott, K.B. (1999), 'Components of processing', in J.K. Foster and M. Jelicic eds., *Memory: Systems, process, or function*. Oxford: Oxford University Press, pp. 32–65.

Rolls, E.T. and Treves, A. (1998), *Neural networks and brain function*, Oxford: Oxford University Press.

Schacter, D.L., Eich, J.E. and Tulving, E. (1878), 'Richard Semon's theory of memory', *Journal of Verbal Learning and Verbal Behavior* 17, pp. 721–743.

Schacter, D.L. and Tulving, E. (1994), 'What are the memory systems of 1994?' in D. L. Schacter and E. Tulving, eds., *Memory systems 1994*, Cambridge, MA: MIT Press, pp. 1–38.

Scoville, W.B. and Milner, B. (1957), 'Loss of recent memory after bilateral hippocampal lesions', *Journal of Neurology, Neurosurgery, and Psychiatry* 20, pp. 11–21.

Skarda, C.A. and Freeman, W.J. (1987), 'How brains make chaos to make sense of the world', *Behavioral and Brain Sciences* 10, pp. 161–195.

Sperling, G. (1960), 'The information available in brief visual presentations', *Psychological Monographs* 74.

Touretzky, D.S. and Redish, A.D. (1996), 'A theory of rodent navigation based on interacting representations of space', *Hippocampus* 6, pp. 247–270.

Tulving, E. (1972), 'Episodic and semantic memory', in E. Tulving and W. Donaldson, eds., *Organization of memory*, New York: Academic, pp. 381–403.

Tulving, E. (1984), 'Multiple learning and memory systems', in K.M.J. Lagerspetz and P. Niemi, eds., *Psychology in the 1990s*, Elsevier, pp. 163–184.

Tulving, E. (1985), 'How many memory systems are there?', *American Psychologist* 40, pp. 385–398.

Tulving, E. (1989). 'Memory: Performance, knowledge, and experience', *European Journal of Cognitive Psychology* 1, pp. 3–26.

Tulving, E. (1999), 'Study of memory: Processes and systems', in J.K. Foster and M. Jelicic, eds., *Memory: Systems, process, or function*, Oxford: Oxford University Press, pp. 11–30.

Tulving, E., Heyman, C.A.G. and MacDonald, C.A. (1991), 'Long-lasting perceptual priming and semantic learning in amnesia. A case experiment, *Journal of Experimental Psychology* 17, pp. 595–617.

Tulving, E., Kapur, S., Craik, F.I., M., Moscovitch, M. and Houle, S. (1994), 'Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography findings', *Proceedings of the National Academy of Sciences (USA)* 91, pp. 2016–2020.

Tulving, E. and Schacter, D.L. (1990), 'Priming and human memory systems', *Science* 247, pp. 301–306.

van Essen, D.C. and Gallant, J.L. (1994), 'Neural mechanisms of form and motion processing in the primate visual system, *Neuron* 13, pp. 1–10.

van Gelder, T. (1995), 'What might cognition be, if not computation', *The Journal of Philosophy* 92, pp. 345–381.

van Gelder, T. (1998), 'The dynamical hypothesis in cognitive science', *Behavioral and Brain Sciences* 21, pp. 615–628.

van Orden, G. C. and Paap, K.R. (1997), 'Functional neural images fail to discover the pieces of the mind in the parts of the brain', *Philosophy of Science* 64, pp. S85–S94.

van Orden, G.C., Pennington, B.F. and Stone, G.O. (in preparation), 'What do double dissociations prove? Inductive methods and isolable systems'.

Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R. and Buckner, R.L. (1998) 'Building memories: Remembering and forgetting of verbal materials as predicted by brain activity', *Science* 281, pp. 1188–1191.

Waugh, N.C. and Norman, D.A. (1965), 'Primary memory', *Psychological Review* 72, pp. 89–104.

Wheeler, M. (in press), 'Explaining the evolved: Homunculi, modules, and internal representation', *Robotics and Autonomous Systems*.

Wimsatt, W. C. (1986) 'Forms of aggregativity', in A Donagan, A.N. Perovich, Jr., and M.V. Wedin, eds, *Human nature and natural knowledge*, Dordrecht: Reidel, pp. 259–291.

Zipser, D. (1985). 'A computational model of hippocampal place fields', *Behavioral Neuroscience* 99, pp. 1006–1018.