To appear in P. Robbins and M. Aydede (Eds.). *Cambridge handbook of situated cognition*. Cambridge: Cambridge University Press.

# Explanation: Mechanism, Modularity, and Situated Cognition[1]

William Bechtel
Department of Philosophy and Interdisciplinary Programs
in Science Studies and Cognitive Science
University of California, San Diego

The situated cognition movement has emerged in recent decades (although it has roots in psychologists working earlier in the 20th century including Vygotsky, Bartlett, and Dewey) largely in reaction to an approach to explaining cognition that tended to ignore the context in which cognitive activities typically occur. Fodor's (1980) account of the research strategy of methodological solipsism, according to which only representational states within the mind are viewed as playing causal roles in producing cognitive activity, is an extreme characterization of this approach. (As Keith Gunderson memorably commented when Fodor first presented this characterization, it amounts to reversing behaviorism by construing the mind as a white box in a black world). Critics as far back as the 1970s and 1980s objected to many experimental paradigms in cognitive psychology as not being ecologically valid; that is, they maintained that the findings only applied to the artificial circumstances created in the laboratory and did not generalize to real world settings (Neisser, 1976; 1987). The situated cognition movement, however, goes much further than demanding ecologically valid experiments—it insists that an agent's cognitive activities are inherently embedded and supported by dynamic interactions with the agent's body and features of its environment.

Sometimes advocates of a situated approach to cognition present their position in an extreme manner that sets the situated approach in opposition to the attempts in cognitive science and cognitive neuroscience to understand the mechanisms within the mind/brain that underlie cognitive performance (Agre, 1995; Beer, 1995; Brooks, 1991; Suchman, 1987; 1993; Thelen & Smith, 1994). Advocates of the *extended mind* perspective maintain that cognitive activities (perceiving, reasoning, problem solving, remembering) do not happen just in the head but extend out into the environment. These environmental factors become, on this view, part of the mind (Clark & Chalmers, 1998). Such challenges are sometimes presented as opposing the search for mechanisms inside the head to explain cognitive activity.

My contention is that an appropriate understanding of the situatedness of cognition does not *require* the denial that the proper locus of control (this notion, developed in Bechtel & Richardson, 1993, will be elaborated upon further below) for cognitive activity is the mind/brain. That is, for *mental* phenomena it is appropriate to treat the mind/brain as the locus of the responsible mechanism and to emphasize the boundary between the mind/brain and the rest of the body and between the cognitive agent and its environment. The phenomena for which such a strategy is not appropriate are ones in which the agent is so intertwined with entities outside itself that the responsible system includes one or more cognitive agents and their environment. These are prototypically social phenomena, not behavioral or psychological ones. There are explanatory principles that determine when it is appropriate to identify the mind-brain as the locus of control and when it is appropriate to identify a larger system as responsible. I will

[1] I thank Carl Craver both for very helpful comments on a draft of this paper and for many productive discussions about mechanism and mechanistic explanation.

articulate a view which maintains that for most explanatory challenges addressed by cognitive science, the mind/brain is the appropriate locus of control even for activities which depend critically on how the agent is situated in an environment.

Reconciling the cognitive science project of identifying and describing the operations of mechanisms inside the head with the claims that cognition is situated requires an appropriate understanding of mechanisms. Mechanisms are bounded systems, but ones that are selectively open to their environment and that often interact with and depend upon their environment in giving rise to the phenomenon for which they are responsible. Moreover, biological mechanisms operate in the context of active, self-maintaining organisms that are dependent on their environments and, yet, are in an important sense autonomous from them. Developing this perspective, however, will require a bit of an excursion through theoretical biology. It will bear fruit as I consider later in the chapter how the mind and its cognitive mechanisms should be considered as distinct systems while also acting through and depending upon the world external to them. Biological mechanisms, including cognitive mechanisms, it will turn out, are always situated and dependent on their environments while also being in a critical sense distinct from them.

The same theoretical considerations that inform discussion of the situatedness of cognition also provide insight into a related controversial topic in cognitive science—the modularity of cognitive systems. In this case, however, the insights into the nature of mechanisms will lead to rejecting the extreme claims made on behalf of modularity. On the surface, advocates of modularity seem simply to be advancing the same claim as those identifying the mind/brain as the locus of control for mental activity, only this time at a finer level, identifying it not with the whole mind/brain, but with a module within it. But this strategy of identifying the locus of control for a mental activity at a finer level in a mental module is, in general, poorly motivated. It fails to consider that explaining the mind/brain's performance of a cognitive task involves decomposing it into component operations, each of which contributes differentially to the performance of the task. Performance of most cognitive tasks requires the orchestrated contribution of many components of the cognitive system, not just one subsystem.

The tension in rejecting modularity and yet treating the mind/brain as the locus of control for cognitive activity should be apparent: Modularity is rejected as failing to recognize the diverse components involved in performing a cognitive task, and advocates of situated cognition likewise maintain that many cognitive activities involve components outside the agent itself. Yet, it is a consistent position to reject modularity within the mind/brain and yet maintain that the mind/brain is the locus of control for cognitive activities. Showing why this is consistent requires developing in more detail the project of explanation in terms of mechanism. To set the context for that, I will begin by exploring the reasoning that has guided advocates of modularity and contrast that with the mechanistic perspective.

## 1. Dividing Minds

Dividing the mind/brain into component systems or modules has been a common strategy in both philosophical and psychological theorizing. Plato's tripartite division of the soul into a reasoning, a spirited, and an appetitive element was an early exemplar. Faculty psychology, as developed in

the 18th century by Christian von Wolff and Thomas Reid, appealed to separate mental faculties responsible for activities such as reasoning, remembering, judging, and willing. At the outset of the 19th century Franz Joseph Gall (1812) aligned the division of the mind into faculties with his differentiation of regions in the brain. Gall's characterization of brain regions in terms of cranial protrusions and indentations was problematic, but the project of localizing mental faculties in the brain obtained greater respectability when Paul Broca (1861) proposed the localization of articulate speech in left prefrontal cortex on the basis of deficits in patients with brain lesions.

The localizationist projects of the 19th century were supplanted in the early 20th century by more holistic views of the brain and the behaviorist tradition in psychology (both traditions are exemplified in the work of Karl Lashley, 1950). The behaviorist tradition emphasized general learning procedures and hence rejected the quest for discrete psychological mechanisms underlying different behaviors. One of the features of the development of cognitive psychology in the 1950s and beyond was the attempt to identify different mechanisms as responsible for different abilities (consider, for example, the different types of memory stores posited in Atkinson & Shiffrin's 1968 classical memory model; as well as the differentiation of memory systems by Tulving and his collaborators, see Schacter & Tulving, 1994).

In this context the term *module* began to be employed for the mechanisms responsible for different types of processing. The term has been used in a variety of ways that emphasize more or less separation of the activities associated with modules. Perhaps the most extreme view of the segregation of modules is found in Fodor (1983). He identifies nine characteristics of modules: they (1) are domain specific, (2) are mandatory in their operation, (3) allow only limited access to the computations of other modules, (4) are fast, (5) are informationally encapsulated, (6) have shallow outputs, (7) are associated with fixed neural architectures, (8) exhibit characteristic and specific breakdown patterns, and (9) exhibit a characteristic pace and sequence in development. Of these, the fifth, informational encapsulation, is both the feature Fodor most emphasizes and the one that makes his account of modules especially strong. Informational encapsulation, for Fodor, entails that a module only employs information encoded within it in its processing; it cannot utilize information stored in another module or in what he terms *central cognition*. Central cognition, in contrast, is holistic in that anything a person knows might be applied in revising one's beliefs or determining how well supported a belief is. As a result of being encapsulated, modules for Fodor do not exhibit much intelligence; accordingly, he views only input processing, language processing, and possibly motor output processing as modular. He construes the modularity of input processing as in fact a virtue. Insofar as input modules cannot be influenced by one's knowledge and expectations, they can provide information about the world that is not theory-laden and can hence provide a theory-neutral basis on which to evaluate competing scientific hypotheses (Fodor, 1984).

Evolutionary psychologists have adopted Fodor's conception of modules without limiting them to input systems. Instead, they "see cognition as modular right through from input to decision processes" (Shettleworth, 2000, p. 54). Although there are weaker notions of modularity available, there is a powerful, if ultimately mistaken, consideration that leads evolutionary psychologists to extend Fodor's conception of encapsulated modules. A major objective of evolutionary psychology is to show how human cognitive abilities such as reasoning about coalitions, detecting cheaters, making risk aversive decisions, or understanding other minds,

could have emerged through evolution. For theorists such as Cosmides and Tooby (1994; see also Cosmides, Tooby, & Barkow, 1992), the evolution of new modules, especially in the relatively recent period since the Pleistocene, is only possible if the modules are encapsulated and able to be selected individually. Cummins and Allen (1998, p. 3) succinctly capture the close affinity evolutionary psychologists identify between modularity and evolution:

> Taking an evolutionary approach to the explanation of cognitive function follows naturally from the growing body of neuroscientific evidence showing that the mind is divisible. . . . The Cartesian view of a seamless whole makes it hard to see how such a whole could have come into being, except perhaps by an act of divine creation. By recognizing the modularity of mind, however, it is possible to see how human mentality might be explained by the gradual accretion of numerous special function pieces of mind.

Not all theorists who invoke mental modules treat encapsulation as the central feature. Daniel Sperber (1994; 2001; 2005), for example, construes as the defining mark of modules that they operate on specific domains of inputs. On such an account, domains such as arithmetic, face-recognition, reading, are viewed as processed by distinct dedicated modules. Even for Sperber, modules operate relatively independently of each other: a module is triggered by input within its specific domain and "once it is performing its function, a module works on its own and is unable to take advantage of information that might be present in the system as a whole but that is found neither in the input nor in the proprietary data-base of the module" (Sperber, 2005).

The opposition to dividing the mind into modules is usually portrayed as stemming from radical holists who emphasize the unity of mind. The rejection of the possibility of dividing minds was a central feature of Descartes contention that the mind could not be a physical entity but must be immaterial. It also figured prominently in Flourens (1824) criticisms of Gall's phrenology and in the early 20$^{th}$ century rejection of neural localization by Lashley (1929) and others. A similar holist bent is manifest in many contemporary dynamical systems theorists who reject the decomposition of the mind into component functions and the attempt to localize such functions in the brain either through lesion experiments or functional neuroimaging (van Orden & Paap, 1997; van Orden, Pennington, & Stone, 2001; Uttal, 2001).

Fodor's rejection of modules in central cognition ironically aligns him (an arch defender of symbolic accounts of cognition) with dynamical systems theorists. His construal of central cognition as utilizing any information the agent possesses reflects a strong holistic perspective. But whereas Fodor sees the inability to differentiate central cognition into modules as undermining the possibility of scientific explanation (Fodor's first law of the non-existence of cognitive science), dynamical systems theory advances a scientific program for explaining the activities of cognitive systems. The strategy is to develop differential equations relating variables that characterize the system being modeled. The nonlinear nature of these equations generates complex patterns of change that can be represented in diagrams showing, for example, the attractor landscape of such a system, but not easily characterized in terms of the behavior of individual system components.

Not only do some dynamical systems theorists resist any attempt to decompose the mind into separate modules (van Gelder, 1995; 1998), they also reject drawing a sharp distinction between

the mind/brain and the rest of the body and the environment in which the mind operates. Equations describing the relations between variables within the brain can be coupled with those characterizing variables external to it (Beer, 2000; Kelso, 1995; Keijzer, 2001). Since on such an approach there is no fundamental difference between variables characterizing the cognitive system and those characterizing features outside the system, the dynamical approach is readily able to integrate phenomena from the mind and the world and capture the embodied and situated nature of cognition.

Although much is often made of the opposition between modular and holistic approaches, I will argue that the dichotomy is actually a false one. This is best appreciated by considering the nature of mechanistic explanation. Whereas most philosophical accounts, including philosophical accounts of psychology, advert to laws as the vehicle of explanation, most explanations advanced in the life sciences make no reference to laws and when laws do appear, they do so in an ancillary role. As Robert Cummins (2000) notes, in psychology laws are typically referred to as *effects* and they typically characterize phenomena in need of explanation but do not themselves explain the phenomenon. Rather, what serves to explain a phenomenon is an account of the mechanism responsible for producing it. As will be discussed below, the parts of biological mechanisms are not totally isolated modules. Rather, the parts of a mechanism are often highly interactive in the production of any phenomenon. Yet, they also have an identity of their own and there are good explanatory reasons to differentiate them from their environmental context.

**Mechanisms and Mechanistic Explanations**

The conception of mechanism has its roots in the machines that humans build. Much of Greek philosophy viewed machines as operating in opposition to nature. For many natural philosophers of the scientific revolution, however, machines came to provide the model for understanding processes in the natural world (Garber, 2002). Descartes extended the idea of mechanism not only to the inorganic world, but to the animate world itself, construing the nervous systems of humans and animals as hydraulic systems in which the flow of animal spirits was altered by sensory experience and directed through the system so as to cause the motion of the limbs. Subsequent to Descartes, the strategy of explaining biological phenomena in terms of machines was pursued by many biologists, although contested by others who insisted that some features of living systems simply could not be explained in mechanical terms and required something extra. This was the basis of the long enduring vitalist/mechanist controversy in biology.

For Descartes and other early mechanists, a mechanism produced its behavior in virtue of the size, shape, and motion of its parts. Over time the repertoire of types of component parts expanded and these parts increasingly were conceived of as entities actively doing things. For example, after Berzelius (1836) introduced the notion of a catalyst as an entity which promoted a chemical reaction without being consumed in it, chemists commonly invoked catalysts to explain reactions that would not otherwise occur under the conditions realized (e.g., at the existing temperature). In the early 19$^{th}$ century many chemists construed yeast not as a living organism but simply as a chemical catalyst that promoted fermentation. Once it was determined to be a living organism in the middle of the century, chemists and subsequently biochemists sought catalysts, later termed *enzymes*, within yeast that could account for fermentation. This project

finally bore fruit in the early decades of the 20th century (Bechtel, 1986; 2006, chapter 3). Implicit in early accounts of mechanism was the fact that the parts of a mechanism had to be appropriately organized in order to perform their functions; the emphasis on organization became far more explicit after Bernard (1865) appealed to the organization in living systems in his attempt to explain how organisms could do the sorts of things vitalists had claimed would not be possible if they were mere mechanisms.

Recently a number of philosophers have advanced accounts of what counts as a mechanism in biology (Bechtel & Richardson, 1993; Glennan, 1996; 2002; Machamer, Darden, & Craver, 2000). My preferred account is:

> A mechanism is a structure performing a function in virtue of its components parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena. (Bechtel & Abrahamsen, 2005; Bechtel, 2006)

Mechanisms exist in nature whereas mechanistic explanation involves an investigator presenting an account of the mechanism taken to be responsible for a given phenomenon. Typically, the explanation involves describing or depicting the component parts, operations, and their organization (diagrams are often far more useful than linguistic descriptions for this purpose). Understanding how the orchestrated operation of the parts produces the phenomenon of interest, investigators must simulate the operation of the mechanism, either mentally or by using model systems or computer simulations.

At the heart of the understanding of a mechanism is the idea that it consists of parts that perform different operations. Already here the conception of components of a mechanism departs from the conception of modules. Although advocates of modules do sometimes talk in terms of submodules, they do not focus on the decomposition of what the overall system does into contributing operations. Rather, a module is identified with a domain of performance—the module performs one of the tasks performed by the overall system. In fact, sometimes the discovery of a mechanism begins in this way (in Bechtel & Richardson, 1993, we spoke of this as simple localization). But over time investigators often learn that other parts are involved in producing the phenomenon and that the part in question only performs one of the required operations. The result is a model of an integrated system responsible for the phenomenon, not a single component.

Identifying the parts and their operations requires decomposing the mechanism. Different techniques enable investigators to decompose a mechanism into component operations (functional decomposition) or into component parts (structural decomposition). Ultimately the goal is to line up the parts with the operations they perform, which Richardson and I refer to as *localization* (Bechtel & Richardson, 1993). Although the notion of levels has proven to be a vexed one (Craver, in press), there is a clear sense in which the parts and operations within a mechanism are situated at a lower level of organization than the mechanism itself. Mechanistic explanation is in this sense reductionistic (Bechtel, in press). The fact that the operations that parts of a mechanism perform are different from the phenomenon exhibited by the whole mechanism and individually do not realize the phenomena makes the working parts of a mechanism different from domain-specific modules.

The parts of a mechanism need not be spatially contiguous but may be distributed throughout the mechanism. What is essential for something to count as a part is that it performs an operation for the mechanism that is different from the operations performed by other parts. The cardiovascular system of an organism and the glycolysis system of a cell each consists of distributed parts that perform different operations. Individual enzymes, for example, catalyze different reactions while individual cofactors are reversibly oxidized in particular reactions. These are distributed through the cytoplasm of the cell but their operations are so coordinated to constitute the mechanism of fermentation. As I will develop in the next section, neither such mechanisms themselves nor their parts are encapsulated from each other in the manner of Fodorian modules.

**Nearly Decomposable Mechanisms versus Encapsulated Modules**

At first pass, mechanistic explanation may not seem to alter the modules/holism dichotomy but rather simply to take the side of modularity. A mechanism is differentiated from its environment and the mechanism is decomposed into parts that perform their own operations. But in fact taking mechanisms seriously and focusing especially on the sorts of mechanisms that arise in biology radically alters the picture. It is no longer appropriate to think in terms of a dichotomy between modular accounts and holistic ones, but a continuum in which the middle is occupied by various designs of mechanisms. The differentiation of mechanisms from one another and division of mechanisms into component parts and operations is partial. This is due to the fact that the operations of the component parts of a mechanism are determined not just by their internal constitution (their subparts, the operations of these subparts, and the way they are organized) but also by both the conditions arising within the mechanism as a result of the operation of other components and external factors impinging on the mechanism. The boundaries of the part partially isolate it from other parts, but do not completely encapsulate it in Fodor's sense. Rather, each part is typically affected in a variety of ways by activity occurring elsewhere in the mechanism.

Assuming decomposability in scientific investigations is a heuristic assumption that is only partially true of any given mechanism. Herbert Simon articulated the idea that natural systems would most likely be nearly decomposable, hierarchical systems and that our ability to understand them depends upon this characteristic. By characterizing natural systems as hierarchical, Simon is claiming that they are "composed of interrelated subsystems, each of the latter being in turn hierarchic in structure until we reach some lowest level of elementary subsystems" (Simon, 1996, p. 184) and moreover that there is a "*small or moderate number*" of types of subsystem at any level in the hierarchy (p. 186). *Decomposability* refers to the independence of the sub-systems at any given level. If the sub-systems are completely independent, except for sending outputs from one sub-system to another, the system is fully decomposable. If the interactions are weak, but not negligible, Simon refers to the system as *nearly decomposable*. Simon offers the following as the main theoretical consequences of near-decomposability:

> (1) in a nearly decomposable system the short-run behavior of each of the component subsystems is approximately independent of the short-run behavior of the other components; (2) in the long run the behavior of any one of the components depends in only an aggregative way on the behavior of the other components (p. 198).

One factor making for nearly-decomposable systems in the organic world is the fact that chemical bonds are of different strengths. Covalent bonds require much more energy to make or break than ionic bonds, and these in turn are stronger than hydrogen bonds. Structures built with stronger bonds will remain stable despite the formation or breaking of weaker bonds, enabling higher-level structures to form without disrupting more basic structures. Thus, the system can be decomposed at one level, leaving in tact structures at levels lower in the hierarchy.

Hierarchical, nearly decomposable systems are not the only possible complex systems. But Simon offers a powerful reason for thinking that natural systems, especially biological systems, will be such—nearly decomposable systems are much more likely to form, especially via processes such as natural selection. Simon illustrates this point through the parable of two watch makers who make equally fine watches of 1000 parts. Tempus shuns hierarchical designs and makes watches in which all 1000 parts must be in the right configuration before the watch is stable. Hora, on the other hand, adopts the principle of hierarchical design in which the whole watch consists of ten stable-subassemblies, each comprised of ten parts that are in turn stable subassemblies made of ten parts. Allowing that there is a small probability (.01) of interruption per addition of a part (for phone calls to take orders, etc.), Simon establishes that it will take Tempus 4,000 times as long to make a watch as Hora. Simon generalizes this point to evolution by arguing that stable subassemblies can be selected for even in the absence of the fully developed system that is responsible for a trait in modern organisms whereas without stable subassemblies the emergence of complex systems would be virtually impossible since it would require a very unlikely constellation of independent events.

Not only does Simon think that systems in nature are nearly decomposable hierarchical ones, but also that we may only be able to understand systems in our world if they are such:

> The fact then that many complex systems have a nearly decomposable, hierarchic structure is a major facilitating factor enabling us to understand, describe, and even "see" such systems and their parts. Or perhaps the proposition should be put the other way round. If there are important systems in the world that are complex without being hierarchic, they may to a considerable extent escape our observation and understanding. Analysis of their behavior would involve such detailed knowledge and calculation of the interactions of their elementary parts that it would be beyond our capacities of memory or computation (p. 207).

Already in Simon's account of Tempus and Hora there is clearly a strong contrast between modular accounts in cognitive science and components of a system for Simon. The parts of a watch do not themselves keep time but perform operations that enable the watch to keep time. Likewise, it is not the tasks carried out by the whole cognitive system that are assigned to parts, but operations which work together to perform the task. Near decomposability allows that the components, to a first approximation, depend only on the overall operation of the other components, not the individual steps in the operation of these other components. Operations within components, Simon maintains, will transpire on a shorter time-scale than operations between components and can be averaged over.

Moreover, modifications in one component that do not affect features on which other components depend can be made independently, allowing selection to promote variants in one component without sacrificing the success of other components.

Although Simon phrases his account in terms of near-decomposability rather than strict decomposability, he does not advance reasons for thinking living systems will be only nearly decomposable rather than strictly decomposable. But a brief glance at the genetics of modern organisms suggests a reason: most genes do not directly code for traits but rather regulate the expression of other genes. This suggests that evolution has not worked simply by promoting individual components, but by modulating the behavior of already existing components. Such modulation of one component by another is not just manifest in genetic regulation, but in the operation of components within organisms. The biochemical system in even relatively simple organisms involves a huge number of interactions between different chemical pathways, allowing different pathways to shunt products elsewhere or recruit materials from elsewhere. This quickly reduces the decomposability of the overall system; yet, it does not yield complete holistic integration. Individual pathways operate semi-autonomously even while coordinating and orchestrating their operation with other components. Scientists can isolate subsystems, either in their models or in experiments, and render the operation of the whole system intelligible in terms of its parts, even while recognizing how different components can also modulate the operation of other components. In their understanding, the components are independent only to *a first approximation*. This first account can then be elaborated in a more refined account that recognizes the interaction of the components.

Simon provides one avenue for appreciating the sorts of differentiation of components that arise in biological systems and how such differentiation differs from that proposed in modular accounts of cognition. But a different perspective can be provided by considering some of the basic demands placed on living organisms, demands which make it important for them to segregate themselves from their environments and ultimately to segregate some of their mechanisms from each other. As a result of being highly organized, biological systems, like humanly constructed mechanisms, must be assembled and will dissipate over time. Unlike humanly constructed mechanisms, however, biological systems cannot rely on external agents either for their initial assembly (development) or for maintenance and repair. The living organism must perform these activities itself.

Performance of these activities requires that organisms exist in energy gradients from which they can extract and utilize free energy and raw materials. Any living system therefore requires metabolic processes that capture and render energy in useable forms. Metabolism is also required to process matter recruited from outside into a form from which its parts can be constructed or reconstructed and additional mechanisms are required to carry out these constructions. While it is conceivable that such processes could occur in an aqueous milieu that imposed no separation from the surrounding environment as long as the requisite metabolites and enzymes were in sufficiently high concentrations, such a set of metabolic and constructive processes would be extremely vulnerable. Biochemical reactions depend upon concentrations of reactants and most

reactions are reversible and will run in the opposite direction when reactant concentrations are unfavorable. Thus, some means must be found to segregate these constituents of living systems from their environments and maintain them in high concentration if these reactions are to function properly. Living systems rely on biological membranes to segregate themselves, and component systems within them, from their environment. Following such a line of reasoning, Tibor Gánti (1975; 2003) incorporated a metabolic system and a membrane construction system as two of the three constituents in his chemoton model of the simplest chemical system exhibiting the features of life. Gánti effectively demonstrated how many features we associate with life would be exhibited by such a system.

Biological membranes are semi-permeable. Even passively they allow some metabolites to pass from the side in which they are in higher concentration to the other in which they are in lower concentration. Accordingly, while segregating many of their constituents from the external milieu, membranes do not cut them off completely. Membranes, moreover, are not limited to passive transport but can incorporate enzymes that actively transport selected substances across them, either to move substances across the membrane that are unable to pass through it on their own or to move substances in opposition to their concentration gradients. Through opportunistically designed transport mechanisms, living systems are able selectively to admit those substances they need to continue the process of constituting and reconstituting themselves and to remove substances, including the waste products of their metabolism, that will prove toxic to their internal operations. Of course these capacities of the membrane do not come for free—they must be paid for in the currency of energy and constructed through the mechanism of catabolism and synthesis. But the critical point is that a minimal living system such as I have characterized constitutes what Alvero Moreno, following Maturana and Varela (1980), calls an *autonomous system*:

> a far-from-equilibrium system that constitutes and maintains itself establishing an organizational identity of its own, a functionally integrated (homeostatic and active) unit based on a set of endergonic-exergonic couplings between internal self-constructing processes, as well as with other processes of interaction with its environment (Ruiz-Mirazo, Peretó, & Moreno, 2004).

A critical feature of an autonomous system is that it is an active system that operates to maintain itself. As such, it imposes a demand on the subsystems (mechanisms) comprising it—their operation is keyed to the survival of the system itself.[2]

Even in this minimal configuration, an autonomous system is operating on its environment by extracting nutrients from it and excreting waste products into it. If the environment is particularly hospitable and constant, such a simple organism may succeed in preserving itself by absorbing metabolites and expelling waste. Many marine invertebrates, such as jellyfish, are osmoconformers—they are isotonic with their salt water environment and rely on that environment to provide the appropriate concentrations of salt and other essential solutes.

[2] I will refer to the whole living organism as a system, not a mechanism per se. The question of whether an organism itself should count as a mechanism is complicated. Typically, one begins the analysis of a mechanism with an account of the phenomenon for which it is responsible. There are a host of different phenomena one might associate with an organism, and depending upon which is selected, researchers will select different parts of the organism as the responsible mechanism and develop a different decomposition of it into parts (Kauffman, 1971).

Typically, however, such a hospitable environment cannot be counted on and an organism must be proactive and generate the right circumstances in its environment or navigate to suitable environments. But the principles already articulated can be extended to allow for a broader range of engagements with the environment, including operations that change conditions in the environment. For example, a cell might excrete chemical substances into the environment that alter the environment in ways advantageous to the organism. Once an organism develops mechanisms for locomotion (e.g., flagella in single-celled organisms), it is no longer dependent upon the environment to bring nutrients to it and remove its waste, but it can move to secure nutrients and avoid the toxic effects of its waste products. These operations of the living system involve changes to the environment outside the organism, but the operations are performed by the organism (or mechanisms within it).

Focusing on the fact that biological systems are autonomous systems in the sense described, we can understand why, in theorizing about them, it is appropriate to construe them as differentiated from their environments. Their autonomy depends upon them having component mechanisms that perform the necessary operations to maintain themselves and that these are organized so as to operate appropriately together. As such, living systems are appropriate objects of analysis. They are what Richardson and I termed *loci of control* for various vital phenomena. This does not mean that they are isolated and totally independent of their environment. Rather, as open systems, they are critically dependent upon their environment for energy and raw materials and to remove their waste products. In more complex organisms, operations reach out into the environment so as to procure these resources. But there is also systemic closure—the parts and operations within the system operate to maintain themselves as a system even as environment conditions change. Understanding how they do so is an important scientific challenge and justifies the strategy of conceptualizing them as independent.

So far I have focused only on segregating the whole living organism from its external environment. Simple living organisms such as bacteria have only a membrane surrounding their whole cytoplasm. This provides for extremely efficient exchange between different chemical constituents such as the metabolic pathways, the process of protein synthesis, and the information storage system (DNA). But as more complex systems evolved, a new problem arose—the possibility that different mechanisms would interfere with one other. A particularly dramatic example is provided by the introduction of hydolytic enzymes that serve to decompose cellular structures as they age or are no longer needed so that their constituents can either be reused by the cell or expelled. Such enzymes play an important role in enhancing the autonomy of these cells. Since such enzymes would clearly be dangerous if they were allowed to float free in the cytoplasm, internal membranes had to evolve to segregate them from the rest. In eukaryote cells several such sets of membranes have evolved to segregate sets of enzymes responsible for different cell operations into distinct organelles. As with the cell membrane itself, though, these membranes are semi-permeable and, while concentrating particular materials, providing them with a hospitable context to perform their operations, and somewhat segregating them from other cell components, they do not impose absolute boundaries. In fact, impenetrable boundaries between component organelles would be extremely deleterious for a cell since coordinating the operations performed by the different organelles typically relies on complex messaging systems linking different mechanisms and parts of mechanisms.

## Situated and Embodied Cognitive Mechanisms

Having argued that mechanistic explanation does not have to endorse either the strongly modular approach of Fodor and evolutionary psychologists or the extreme holism of dynamicist critics, but allows for identifying systems on the a continuum between them, I turn now to the specific implications for thinking about cognition as embodied and situated. As we will see, the account I have offered supports an understanding of embodiment and situatedness without requiring that we extend the mind out into the world or deny the differentiation of the mind/brain from the rest of the organism and the external world.

We have already seen in the previous section that living organisms are differentiated from their environments as systems that construct and reconstruct themselves. Moreover, in order to understand how organisms accomplish this, scientists need to differentiate the organism from its environment while recognizing the various ways in which the organism engages its environment. In the account so far I have focused only on individual cells and single-celled organisms. The evolutionary route from single-celled organisms to multi-celled organisms is far from clear, but it is evident that true multi-celled organisms are more than an aggregation of single-celled organisms. Rather, they involve a differentiation of cell types that perform different operations. Once cells are differentiated and perform different functions, some means of integrating them into an operative whole is required. Such integration does not obviate the demand that individual cells maintain themselves, but extends the resources for doing so by allowing for specialization of the functioning of individual cells so that each performs a different set of operations needed by the others. Accordingly, theorists such as Rudolf Virchow (1858), who played a major role in establishing that cells derived from pre-existing cells via cell division, conceived of multi-celled organisms as cell republics.[3] Although each cell is a living unit, division of tasks becomes possible, making each dependent upon the operations of the others. As a result, a cell republic (multi-celled organism), like a political republic, is capable to doing things that an individual cell or person cannot. The whole organism now becomes an autonomous system, needing to construct and maintain itself in the face of environmental factors that would lead to its dissipation.

The division of labor and mutual dependency between parts of the system is especially clear with organisms comprised of different organ systems. Individual organs perform different operations that are required by the whole organism—extracting nutrients and oxygen from the environment, distributing these through the organism, executing locomotion, etc. Segregating these activities in different organs allows each to perform its operations without continual interference from the others. But it is also important that these systems remain open to each other so as to maintain the appropriate conditions for the operation of each component. In an early attempt to understand such coordinated operation, and thereby provide an answer to vitalists (such as Bichat, 1805) who thought maintaining life was beyond the capability of any mechanism, Claude Bernard (1865) introduced the idea of differentiating two environments—what is normally termed the environment in which the organism as a whole lives and the *internal environment* in which the

[3] Virchow's emphasis in developing this metaphor was to focus the study of living processes, especially those involved in pathology, on the individual cells. Cells were the most basic organized form—"the last constant link in the great chain of mutually subordinated formations that form tissues, organs, systems, the individual. Below them is nothing but change."

different organ systems operate. By construing each organ as responsive to the conditions of the internal, not external, environment, Bernard proposed to show that their operations were causally determinate, which Bichat had denied. But more importantly, he viewed each organ as operating so as to maintain specific aspects of the internal environment in a constant condition, thereby making the whole organism stable against perturbations in the external environment. This idea was further developed by Walter Cannon (1929), who introduced the concept *homeostasis* and described a number of mechanisms through which organs of the body helped maintain homeostasis. These mechanisms may involve behaviors of the whole organism that configure the environment in a manner that preserves the homeostasis of the individual (Richter, 1942-1943).[4]

Many of the control mechanisms Cannon identified involved the brain, specifically the autonomic nervous system. As important as the brain is as a regulator of what occurs elsewhere in the body, there is a risk of focusing exclusively on it. When we conceptualize control, we often think hierarchically and situate all decision making at the top of the hierarchy. This, however, works poorly both in biology and social institutions. As a result, biological systems usually have multiple layers of control arranged such that higher-level control systems can bias the functioning of lower-level ones (often by affecting the conditions under which more local control systems operate), but do not directly determine the behavior of the lower level systems. This can be appreciated by focusing on organisms in which cortical level control systems have been removed—in such cases many functions continue unimpaired but cannot be (directly) coordinated in the service of higher level objectives (Stein & Smith, 1997). (They can sometimes be indirectly coordinated via their interaction with other components of the system that may still be under higher-level control. Thus, in patients in which a severed corpus callosum prevents direct communication between the two hemispheres of the brain, one hemisphere can still learn from observing the resulting behavior what motor commands the other hemisphere has generated.) The existence of multiple control systems all modulating the behavior of local components requires that these components not be encapsulated from other components of the system but open in appropriate ways to them.

The perspective on organisms as autonomous systems maintaining themselves through their activities, including activities that modify the world around them, provides a way both to conceptualize the mind/brain as an organ of a living organism and as embodied and situated in the world. As part of an organism, it is differentiated from the world in which it operates but is nonetheless highly connected to that world. Even the simplest living organisms, as we have already noted, are distinct autonomous systems that extract energy and raw materials from their environment and put these to use to constitute and reconstitute themselves. But in more complex animals, this will involve performing a larger variety of behaviors (predating on other organisms or avoiding predators) and navigating the environment. These interactions with the environment alter it, often in ways that impact on the organism itself. Accordingly, there is both isolation from the environment as the organism maintains its own identity and engagement with it.

The dependencies on the environment are particularly important in understanding higher cognitive tasks. There has been a tendency to think of these as occurring solely in the organism. This runs the risk of assuming the mind can do more than it in fact can. The symbolic tradition in

[4] I thank Carl Craver for pointing me to Richter's development of homeostasis to include behaviors that served to maintain internal homeostasis.

cognitive science, for example, tends to assume that the mind itself has the power of a universal computer. In it useful in this regard to recall how Turing (1936; see also Post, 1936) himself was led to the conception of a Turing machine as an abstract model for a computer. Extant computers—humans whose profession was to carry out complex arithmetic calculations—provided the model for the Turing machine. These individuals learned and applied a finite number of procedures to problems which were written on paper. In turn they wrote the results of successive operations on the paper and used these as inputs for further operation. The human provided the model of the finite state device in the Turing machine while the paper became the model for the tape (or memory in the computer). In thinking of the mind itself as a Turing machine or a computer, the operations that reached out into the environment are resituated inside the mind. But the mind/brain may not have such resources and by thinking it does, cognitive scientists may set themselves up for failure.[5]

Nonetheless, while relying on environmental resources, it is still the cognitive agent that is performing these activities in pursuit of its ends. It is the cognitive agent that has interest in performing the task and in recruiting components of its environment to enable such performance. Indeed, in the case of the human computer performing calculations, he or she was typically doing so for basic biological ends—securing a paycheck that would provide the food and other resources needed to maintain himself or herself. In this the cognitive agent is like autonomous biological systems that perform operations in their environment so as to secure matter and energy needed to build and repair themselves and dispose of wastes that are toxic to themselves.

Just as the move from individual cells to cell republics resulted in identifying a new locus of control at the level of the organism for the behaviors of the organism, so researchers may find the need to move to a more inclusive system for explaining some phenomena involving humans or other animals. Just as individual cells may specialize their operations and coordinate them so as to maintain a multi-celled organism, so individual organisms may specialize their activities and coordinate them to maintain a larger systems such as a social network. In these cases, the social network becomes the locus of control for certain phenomena—those that are carried out by the social network in the service of it. Such activity, however, takes us beyond situated cognition to social activity. Moreover, there is a principled reason for shifting the locus of control to the social network: it is the network itself that is being maintained by the operations being performed (either between the constituents of the network or in the environment in which the social network is situated). Situated cognition, though, refers to the cognitive activities of agents situated in an environment, and the locus of control for these cognitive activities remains the individual cognitive agent.

## Conclusions

By considering the kind of explanation appropriate to biological systems, mechanistic explanation as applied to biological organisms, I have extracted some insights into both the modularity of the mind/brain and the situatedness of cognition. Biological systems are typically

[5] Turing's initial perspective was rekindled in Rumelhart, Smolensky, McClelland, & Hinton's (1986) account of human performance in multiplying large numbers. Such activity may rely not on internal symbols but external ones on which a mind, operating like a connectionist network that has learned to associate one pattern with another, may operate. For further discussion, see Clark (1987) and Bechtel and Abrahamsen (2002).

bounded and there are good reasons for the mechanisms within them, including cognitive ones, to be segregated from each other. But the boundaries between the organism and its environment and between components and subcomponents within it are permeable. Accordingly, even when we identify a particular system as a locus of control of a particular function, we need not impute full responsibility to that component. Its operation may be dependent upon features of its environment, whether an internal environment within the organism or an external environment in which the organism functions. It may act on and alter its environment in ways that facilitate maintaining itself as an individual system. Thus, we can demarcate the cognitive system while still examining how it is situated in and interactive with the rest of the organism and the environment in which the organism is situated.

Turning within organisms, I have identified reasons why it is useful to segregate different operations in different organs or parts of the system. Unlike in appeals to modularity, the focus in developing mechanistic explanations is on decomposing the overall activity into component activities. Segregating, however, does not mean isolating. In fact, living systems are typically highly integrated despite the differentiation of operations between different organs and cell types. The mind brain seems to be no different on this score—it consists of component processing areas that perform different computations which are nonetheless highly integrated with each other. Such a mechanism does not typically include encapsulated modules, and one is not likely to find them in the mind/brain.

Turing to the whole organism, the traditional view, which treats the skin as the boundary of the organism and the mind as coterminous with the brain and central nervous system, is well-motivated. The organism is the system that maintains itself as a result of the operations it performs, and the brain and central nervous system comprises the system within it that performs critical regulatory tasks. The mind-brain itself and the organism as a whole are open systems and dependent upon the environment; hence, the quest to understand how a cognitive agent together with its various cognitive mechanisms is situated in its environment is also well motivated.

## References

Agre, P. E. (1995). Computational research on interaction and agency. *Artificial Intelligence, 72*, 1-52.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & 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. (1986). The nature of scientific integration. In W. Bechtel (Ed.), *Integrating scientific disciplines* (pp. 3-52). Dordrecht: Martinus Nijhoff.

Bechtel, W. (2006). *Discovering cell mechanisms: The creation of modern cell biology*. Cambridge: Cambridge University Press.

Bechtel, W. (in press). Reducing psychology while maintaining its autonomy via mechanistic explanation. In M. Schouten & H. Looren de Jong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience and reduction*. Oxford: Basil Blackwell.

Bechtel, W., & Abrahamsen, A. (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks* (Second ed.). Oxford: Blackwell.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421-441.

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

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence, 72*, 173-215.

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

Bernard, C. (1865). *An introduction to the study of experimental medicine*. New York: Dover.

Berzelius, J. J. (1836). Einige Ideen über bei der Bildung organischer Verbindungen in der lebenden Naturwirksame, aber bisher nicht bemerke Kraft. *Jahres-Berkcht über die Fortschritte der Chemie, 15*, 237-245.

Bichat, X. (1805). *Recherches Physiologiques sur la Vie et la Mort* (3rd ed.). Paris: Machant.

Broca, P. (1861). Remarques sur le siége de la faculté du langage articulé, suivies d'une observation d'aphemie (perte de la parole). *Bulletin de la Société Anatomique, 6*, 343-357.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139-159.

Cannon, W. B. (1929). Organization of physiological homeostasis. *Physiological Reviews, 9*, 399-431.

Clark, A. (1987). *Microcognition: Philosophy, cognitive science, and parallel distributed processing.* Cambridge: MIT Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7-19.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. S. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind* (pp. 85-116). Cambridge: Cambridge University Press.

Cosmides, L., Tooby, J., & Barkow, J. H. (1992). Introduction: Evolutionary psychology and conceptual integration. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 3-18). New York: Oxford.

Craver, C. (in press). *Explaining the brain: What a science of the mind-brain could be*. New York: Oxford University Press.

Cummins, D. D., & Allen, C. (1998). Introduction. In D. D. Cummins & C. Allen (Eds.), *The evolution of mind* (pp. 3-8). Oxford: Oxford University Press.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-144). Cambridge, MA: MIT Press.

Flourens, J. P. M. (1824). *Researches Expérimentales sur les Propriétés et les Fonctions du Système Nerveux dans les Animaux Vertébris*. Paris: Crevot.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *The Behavioral and Brain Sciences, 3*, 63-109.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1984). Observation reconsidered. *Philosophy of Science, 51*, 23-43.

Gall, F. J. (1812). *Anatomie et physiologie du systême nerveaux et général, et du cerveau en particulier, avec des observations sur la possibilité de reconnoitre plusieurs dispositions intellectuelles et morales de l'homme et des animaux, par la configuration de leur têtes*. Paris: F. Schoell.

Gánti, T. (1975). Organization of chemical reactions into dividing and metabolizing units: The chemotons. *BioSystems, 7*, 15-21.

Gánti, T. (2003). *The principles of life*. New York: Oxford.

Garber, D. (2002). Descartes, mechanics, and the mechanical philosophy. *Midwest Studies in Philosophy, 26*, 185-204.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis, 44*, 50-71.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*, S342-S353.

Kauffman, S. (1971). Articulation of parts explanations in biology and the rational serach for them. In R. C. Bluck & R. S. Cohen (Eds.), *PSA 1970* (pp. 257-272). Dordrecht: Reidel.

Keijzer, F. (2001). *Representation and behavior*. Cambridge, MA: MIT Press.

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

Lashley, K. S. (1929). *Brain mechanisms and intelligence*. Chicago: University of Chicago Press.

Lashley, K. S. (1950). In search of the engram. *Symposia of the Society for Experimental Biology, IV. Physiological Mechanisms in Animal Behaviour*, 454-482.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Maturana, H. R., & Varela, F. J. (1980). Autopoiesis: The organization of the living. In H. R. Maturana & F. J. Varela (Eds.), *Autopoiesis and Cognition: The Realization of the Living* (pp. 59-138). Dordrecht: D. Reidel.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco: W. H. Freeman.

Neisser, U. (1987). Introduction: The ecological and intellectual bases of categorization. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 1-10). Cambridge: Cambridge University Press.

Post, E. L. (1936). Finite combinatorial processes - Formulation I. *Journal of Symbolic Logic, 1*, 103-105.

Richter, C. P. (1942-1943). Total self-regulatory functions in animals and human beings. *Harvey Lectures, 37*, 63-103.

Ruiz-Mirazo, K., Peretó, J., & Moreno, A. (2004). A universal definition of life: Autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere, 34*, 323-346.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemas and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart & T. P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 1-38). Cambridge, MA: MIT Press.

Shettleworth, S. (2000). Modularity and the evolution of cognition. In C. Heyes & L. Huber (Eds.), *The evolution of cognition* (pp. 43-60). Cambridge, MA: MIT Press.

Simon, H. A. (1996). *The sciences of the artificial* (Third ed.). Cambridge, MA: MIT Press.

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39-67). Cambridge: Cambridge University Press.

Sperber, D. (2001). In defense of massive modularity. In E. Dupoux (Ed.), *Language, brain, and cognitive development: Essays in honor of Jacques Mehler*. Cambridge, MA: MIT Press.

Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind: Structure and content*. Oxford: Oxford University Press.

Stein, P. S. G., & Smith, J. L. (1997). Neural and biomechanical control strategies for different forms of vertebrate hindlimb motor tasks. In P. S. G. Stein, S. Grillner, A. I. Selverston & D. G. Stuart (Eds.), *Neurons, networks, and motor behavior* (pp. 61-73). Cambridge, MA: MIT Press.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge: Cambridge University Press.

Suchman, L. (1993). Response to Vera and Simon's situated action: A symbolic interpretation. *Cognitive Science, 17*, 71-75.

Thelen, E., & Smith, L. (1994). *A dynamical systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, second series, 42*, 230-265.

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.

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

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21*, 615-628.

van Orden, G. C., & 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*(4), S85-S94.

van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? Inductive methods and isolable systems. *Cognitive Science, 25*, 111-172.

Virchow, R. (1858). *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre*. Berlin: August Hirschwald.