To appear in *Bulletin d'Histoire et d'Epistémologie des Sciences de la Vie*

Mechanistic Explanation and the Nature-Nurture Controversy

William Bechtel and Adele Abrahamsen
University of California, San Diego

Both in biology and psychology there has been a tendency on the part of many investigators to focus solely on the mature organism and ignore development. There are many reasons for this, but an important one is that the explanatory framework often invoked in the life sciences for understanding a given phenomenon, according to which explanation consists in identifying the mechanism that produces that phenomenon, both makes it possible to side-step the development issue and to provide inadequate resources for actually explaining development. When biologists and psychologists do take up the question of development, they find themselves confronted with two polarizing positions of nativism and empiricism. However, the mechanistic framework, insofar as it emphasizes organization and recognizes the potential for self-organization, does in fact provide the resources for an account of development which avoids the nativism-empiricism dichotomy.

Mechanistic Explanation

The received philosophical view of explanation is one which involves invocation of a law and a demonstration that given the law and initial conditions, the phenomenon to be explained is what is expected (Hempel & Oppenheim, 1948; Hempel, 1966). When one looks into recent literature in either biology or psychology, however, it is striking how infrequently laws are invoked. Instead, explanatory activity tends to be directed towards understanding the mechanism that produces the phenomenon of interest. Mechanistic explanation is often incomplete or programmatic, as when it takes the form of generalizations that point to design features of a mechanism or identify variables that influence its performance. However, especially in biology, mechanistic explanation also can take the form of explicit models that become increasingly detailed as research fills in gaps and resolves puzzles. Given the prominent role of conceptions of mechanism in the life sciences, and their relative neglect in philosophy of science until recently, we begin by asking: just what is a mechanism?

The appeal to mechanisms figured prominently in the scientific revolution, with natural philosophers such as Descartes and Boyle championing what Boyle referred to as the *mechanical philosophy*. For Descartes, this involved appeal to the shape and motion of component parts (with minute corpuscles comprising the ultimate components out of which objects in the physical world were constructed and from which they derived their properties). As influential as it was, Descartes' conception of a mechanism turned out to be inadequate to explain either psychological phenomena (as Descartes recognized in making his case for dualism) or biological phenomena (as the criticisms of Xavier Bichat and other vitalists made clear). Beginning with the work of Claude Bernard in the 1860s, biologists came to emphasize a far more complex conception of mechanism which made critical reference not only to components but also to the particular ways in which they are organized into cohesive systems. Although the conception of mechanism has changed over time, the pursuit of mechanistic explanations of phenomena has remained a staple of both biology and psychology.

Within the past decade a number of philosophers (Bechtel & Richardson, 1993; Glennan, 1996; Machamer, Darden, & Craver, 2000) have sought to capture the essential features of mechanisms as they appear in explanations in the life sciences. Although there are minor differences between these accounts, a common core is that a mechanism is construed as a composite system that performs an activity as a result of being comprised of components, each performing a particular operation, whose organization ensures that the operations are coordinated so as to perform the activity (for an account that applies this framework specifically to psychology, see Wright & Bechtel, in press). We will highlight a few of the central features of this conception of mechanism that are relevant to the issues concerning development. First, mechanisms are characterized in terms of the activities they perform. Within the same region of space-time, different activities may be occurring, and what counts as comprising a mechanism will depend upon what activity an investigator is seeking to explain. Second, explaining an activity performed in a given context requires decomposing that activity into component operations and localizing them—that is, linking them to component parts of the system. Third, critical to the explanation is determining the organization within the system that allows the component operations to be coordinated appropriately to produce the overall activity. Fourth, the operation of a component can itself be treated as an activity to be explained by another round of decomposition and localization at a lower level; repeatedly carrying out this process generates a cascade of finer and finer-grained mechanistic accounts. (The willingness to repeatedly decompose a system into parts and to localize operations in those parts makes mechanistic explanation a reductionist program. But the emphasis on organization, and the recognition that the activity of a mechanism is influenced by its environment, including any larger-scale mechanism in which it is incorporated, add an appreciation for the roles played by higher levels of organization that typically is missing from philosophical treatments of reductionism.)

The challenge for explaining development from a mechanistic perspective already appears in this sketch of what a mechanism is. Although the components of a mechanism must do things and interact with each other in coordinated ways in order for the mechanism to perform its activity, these components are viewed as givens—that is, as enduring entities that are unchanged by their history of performing the same operations again and again. Further, there is an implicit assumption that there is in place a particular organization of the components which also is unchanged by the system's history. A mechanism may undergo transitory changes, but these changes are restricted by its seemingly unchanging organization and limited set of operations. The same changes occur again and again in what is the same mechanism. But how did the mechanism arise in the first place? How did it develop into the sort of mechanism that it is?

There is an important clue already in the conception of a mechanism for addressing the question of how mechanisms develop. This is the fact that central to the operation of a mechanism is organization. A mechanism is not just a set of components doing things, but an organized set of components. The organization is often critical in determining what the mechanism does. It is in virtue of being organized that a mechanism does things that its components alone cannot do. This is especially true when a mechanism has an appropriate nonlinear organization—that is, the operation of at least some of the components feeds back to affect the operation of other components (or of themselves). This kind of organization can bring about nuanced, effective self-regulation. If we focus on how such organization might become established, and especially on the capacity of various kinds of components to self-organize and to maintain and alter their own organization, we do have an important clue as to how we might provide an account of the development of mechanisms.

We will return to this clue in a later section. For now, what is important to note is that a focus on how mechanisms operate in mature systems (e.g., adult humans) tends to push the question of development out of consideration. Investigators take mechanisms as already in place and consider only how they work in response to inputs from the environment. This tendency to ignore development is compounded by a perception that if one does address questions of development, a choice must be made between two unacceptably polarized options: that the mechanism essentially pre-existed or that it came into being over time. In the next section, we will explore what has become of these polarizing positions.

Polarizing Positions

Within both biology and psychology, when questions of development have been raised, the tendency has been to push positions that appeal primarily to nature (innate endowment) or to nurture (changes over time in which internal and/or external environment play the major causal role). In biology a classic nature position was preformationism (Malpighi, 1675), which held that an organism was already preformed in the egg—in the most concrete versions, as a kind of miniature adult that needed only to grow. This was opposed by those advocating epigenesis, according to which initially undifferentiated material gradually differentiates into the structures of the mature organism (Harvey, 1651; Wolff, 1759). Causation was not well-addressed in classic epigenesis, but involved some outside force or tendency acting on the undifferentiated material in the egg. In the stark terms in which the alternatives were initially posed, the controversy ended in the 19th century with the general claim of victory for epigenesis. But the controversy was not laid to rest—it simply took on new forms. In its contemporary guise, the preformationist position is now embodied in the view that development is strictly under genetic control and consists in constructing the organism whose plan was laid down in the genome. The epigenetic position is now particularly a concern of developmental biologists, who show considerable variation in their emphasis. Closest to the nurture pole are those who emphasize the cellular environment and regard genes more as an influence on events in that environment than as a plan to be realized. Other developmental biologists (as well as developmental systems theorists) have played an important role in establishing a middle ground by viewing the problem as one of determining the often complex ways in which genes interact with their cellular environment to produce development.

What kind of system is needed to realize a middle ground is the focus of this paper, but we will address this in the context of psychology and the cognitive sciences rather than biology. In these disciplines, the polarized alternatives derive from the 17th and 18th century epistemological traditions of rationalism and empiricism. Rationalism, as it developed in the theories of Descartes, Leibniz, and others, emphasized the capacity of reason, properly employed, to establish truths. Reason in these accounts operates by applying principles of logic to indubitable premises. For example, Descartes starts with the requirement of clear and distinct ideas, which provide certain knowledge and from which reason can derive other truths. For rationalists to explain how the results of such reasoning could be true about features of the external world, they had to invoke such principles as the claim that God was not a deceiver and so would not provide us with false initial ideas or corrupt procedures for making inferences from these starting points. From a contemporary point of view, the crucial claim of rationalism is that the project of acquiring knowledge begins with powerful innate ideas. This was rejected by the empiricist tradition, which insisted that information about things in the world must be developed through sensory experiences ("there is nothing in the mind that is not first in the senses"—an idea that is found in both Aristotle and Thomas, predating the empiricism that arose in the 17th century). Such empiricists as Locke and Hume proposed accounts based on principles of association whereby ideas derived from sensory impressions were built up into general understanding. This tradition was further developed by David Hartley in the 18th century, by John Stuart Mill and Alexander Bain in the 19th century, and by the behaviorists in the United States in the first half of the 20th century. The principles of association, it was recognized, must be provided by the mind. Thus, empiricists did not maintain that nothing was innate, but insisted that only very general procedures for learning were innate, not any particular knowledge.

This longstanding conflict emerged with renewed vigor when Noam Chomsky made it a major focus of his reconstrual of linguistics. In uncompromising opposition to behaviorist accounts of language learning (especially Skinner's), he argued that knowledge of grammar must be innate (Chomsky, 1959). For Chomsky, learning a native language does not involve learning the fundamentals of grammar, but rather determining the specific implementation of those fundamentals that is employed in one's native language. (In his 1980s principles and parameters formulation, for example, Chomsky proposed that the appropriate value of each innate parameter was selected through encounters with sentences in the language to be learned.) A key part of his argument was that language learners are not provided with input that is adequate for learning a grammar from scratch. Chomsky referred to this as the poverty of the stimulus argument (Chomsky, 1980, p. 34). Just as a theory goes beyond particular facts to allow the prediction of additional facts, a grammar goes beyond the actual utterances one hears to predict, in principle, all of the acceptable sentences of a language. Despite the fact that the stimulus from which a child must learn her language is impoverished relative to the grammar of the language, children typically acquire it. Given the limited input children receive, from an empiricist perspective one might think they would make a variety of incorrect inductions—for example, that a question is formed by fronting the auxiliary of the first clause ("Is the man who tall is in the room?") rather than that of the main clause ("Is the man who is tall in the room?"). This is because the first clause is so often also the main clause, providing the child with little evidence as to which induction would be the correct one. From Chomsky's rationalist perspective, the absence of such errors is explained as resulting from innate knowledge of grammar. Chomsky's arguments are frequently buttressed by appeal to formal proofs developed by Mark Gold (1967) showing that it is impossible to induce the correct grammar of a language wholly from positive examples. The space of possible grammars is too great and must be constrained.

As a complement to Chomsky's arguments, Fodor (1975) offered arguments that mental representations are innate. He contended that it was impossible to acquire the concepts of a language via a process of hypothesis testing, because the hypothesis about the meaning of the concept already had to be represented before evidence for or against that hypothesis could be evaluated. But if one could already represent the meaning, then one already possessed the concept. All the hypothesis could do would be to propose a connection between a new representation and the already existing one. Thus, Fodor contended that language learners required an innate *language of thought*, sometimes referred to as *mentalese*.

One of the key claims Chomsky made in developing his nativist account of language is that linguistic knowledge resides in a specialized language faculty—a language organ—analogous to bodily organs such as the liver or heart. Fodor (1983) reinterpreted this idea and took it much further by developing and promoting his modularity thesis. With language as his primary but not only example of a mental module, Fodor suggested a number of criteria: modules (1) are domain-specific, (2) operate in a mandatory fashion, (3) are fast, (4) are informationally encapsulated, (5) have shallow outputs, (6) are associated with fixed neural architecture, (7) exhibit characteristic and specific breakdowns, and (9) manifest a characteristic developmental pattern. In addition to the language module, Fodor identified modules in the initial processing of sensory inputs, citing, for example, the inability of subjects to change the way they see visual illusions as evidence that these systems are encapsulated processing systems (the feature he ultimately construed as the one most diagnostic of modularity). In contrast to input processing, Fodor maintained that central cognition, in which a person could bring any knowledge to bear on a given problem, was not modular. (In a pessimistic and perhaps mischievous vein, though, he proposed in his first law of the non-existence of cognitive science that central cognition was not amenable to empirical analysis.)

Although Chomsky focused on language, and Fodor extended that focus only so far as to include other input systems, the nativist perspective recently has been expanded to other mental capacities by a research program that has adopted the name *evolutionary psychology* (Barkow, Cosmides, & Tooby, 1992; Cosmides & Tooby, 1994; Sperber, 1994). This program seeks to analyze the mind into a number of separate modules that perform specific activities that facilitate an organism's reproductive success, for example, the detection of cheaters. Each of these modules is then assumed to have an evolutionary history and to exist in current organisms because it contributed to the fitness of their ancestors. (For a discussion of the neural implausibility of such modules, see Bechtel, 2003.)

A Third Contender?

In psychology and the cognitive sciences, as in biology, numerous investigators have sought a middle ground between the polarized positions of nativism and empiricism. Almost invariably, such positions are rejected by advocates of nativism as essentially empiricist, but distinctive and powerful ideas have emerged from such efforts. A prime example is Piaget's account of development. His position is sometimes called *interactionism* because he emphasized the interaction of nature and nurture during development, and is also known as *constructivism* because he viewed mental structures as being constructed and becoming more complex in the course of children's active engagement with their world. It is Piaget's constructivism that is the key idea here, because without a way to build more complex structures from simpler structures, the interaction between nature and nurture will not yield interesting outcomes. Piaget was limited to the tools available in his era, but in recent decades an approach commonly referred to as *connectionism* or *neural-network modeling* has given new life to these ideas. We will briefly introduce this approach and then illustrate how it can provide very explicit exemplars of, first, interactionism and, second, constructivism.

For connectionists, cognitive processes are realized in the activity of networks of simple neuron-like units. In the simplest arrangement, each unit in an input layer has a weighted connection to each unit in an output layer. An input to the network is imposed as a pattern of activation across the input units, and the network transforms this into an output pattern by applying an activation function that is sensitive to the weights. Similar models had been advanced in the 1940s through the 1960s (Rosenblatt, 1962), but powerful arguments regarding intrinsic limitations in the capacity of such systems (Minsky & Papert, 1969) contributed to the temporary demise of this tradition. Its reemergence in the 1980s was to a large degree the result of the promulgation of a powerful new learning algorithm, *backpropagation* (Rumelhart, Hinton, & Williams, 1986), which permitted the learning of input-output mappings that had been beyond the capacity of earlier networks. If a network had at least one intermediate layer of *hidden units* between the input and output layers and a non-linear activation function, backpropagation could be used to gradually change its weights during repeated exposure to a training corpus (pairs of input-output patterns). Networks with hidden layers had been known to Rosenblatt, but he did not have a learning algorithm that would guarantee finding appropriate weights on the connections between units for a given training corpus. Eventually it was established that networks using the backpropagation learning algorithm, as long as they had sufficient hidden units, could be trained to perform any input-output mapping.

Nativists generally have regarded models of this kind as excessively rooted in the empiricist-associationist tradition and unable, in principle, to capture the sorts of knowledge they regard as innate. One criticism is that a network, prior to training, is essentially a Lockean *tabula rasa* that provides no means of representing innate knowledge. Another criticism is that even after training, a network is simply a set of associations in which the degree of association is indexed by connection weights, whereas many kinds of knowledge must be specified in rules involving variables, as exemplified in generative grammars or traditional artificial intelligence programs (Fodor & Pylyshyn, 1988; Fodor & McLaughlin, 1990). The second criticism generated considerable attention and heated debate in the 1980s as to whether rules are necessary (for a discussion of this as well as a more general discussion of connectionism, see Bechtel & Abrahamsen, 2002). But it is the first criticism that is most relevant here.

The first criticism—that connectionist models are instruments of empiricism that cannot represent innate knowledge—generated relatively little debate because many connectionists had strong empiricist leanings and saw nothing to contest. We maintain, instead, that one of the most promising roles for connectionist modeling is to serve as an explicit medium for pursuing an interactionist/constructivist perspective. That raises the issue of whether, or how, sufficient innate endowment for fruitful interaction can be incorporated in networks. A major leap forward in addressing this issue was the publication of a book, *Rethinking Innateness*, which was the product of collaboration among several constructivist connectionists (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996). One of their contributions was to distinguish three classes of innate constraints. The class that applies most broadly to connectionist models is architectural constraints: whether the network is layered, the number of layers, the number of units in each layer, the activation function used, and so forth. Regardless of whether or not a network designer thinks of decisions like these as providing the network with an innate endowment, they do function in that way—they influence what the network can learn (or can learn easily). Some simulations have been especially informative about the consequences of providing networks with different native endowments via architecture. For example, connection patterns can be restricted such that sets of relatively segregated units interconnect primarily with each other to form subnetworks. (These are called modules, but in a somewhat weaker sense than Fodor's.) Jacobs, Jordan and Barto (1991) designed a modular network in which the modules differed architecturally—one had no hidden layer. When they trained the network on two distinct tasks, each module "decided" to specialize on the task most suited to its particular architecture.

Though it is important to explore such effects, situations in which architecture is not destiny are equally interesting. The simulation of some important developmental phenomena is robust across quite different architectures. In certain linguistic and cognitive domains, for example, children exhibit U-shaped learning curves rather than steady improvement. In the domain of inflectional morphology, preschoolers (1) master the correct past-tense form of a few irregular verbs (e.g., *ran* as the past-tense of *run*); (2) later overgeneralize the regular past-tense to some of these irregular verbs (e.g., they may use *runned* as the past-tense of *run*); and (3) eventually re-exhibit the correct form. A connectionist explanation of this developmental sequence was first provided by a simulation involving a large, nonlayered network that used what connectionists call a "distributed" representation of verb forms (Rumelhart & McClelland, 1986). U-shaped learning later was demonstrated as well in simulations involving a small, layered network that used a "localist" representation of verb forms (Plunkett & Marchman, 1991).

Elman et al.'s second class of constraints involves the timing of developmental events. Timing can be imposed by the environment (e.g., presenting some parts of a corpus to a network before other parts), by a maturational process (e.g., new units are added on a predetermined schedule), or have more complex determinations (e.g., new units are added when a particular level of learning has been achieved). Some of the most interesting connectionist research involves the incorporation of timing constraints, as illustrated in the simulation by Elman, 1991, in the next section.

The third class of constraints is the one most relevant to classic, polar nativism and least explored in connectionist modeling: innate representations (i.e., innate knowledge or content). In a connectionist network, knowledge is not represented in symbol strings that can be directly accessed or "read off"; rather, it is distributed through the network in the form of weights on connections. (Input and output patterns often look more like classic representations, but they are fleeting states of the network that result from its interaction with the environment and are not permanently represented.) In most connectionist simulations, the initial weights on the connections are random, and each task is learned de novo. This shows off the networks' learning capabilities, but there is no reason in principle that networks could not begin with weights more suited to the task they were to learn, and thus incorporate innate knowledge in this third, strong sense.

There is considerable disagreement about what kinds of constraints are actually innate in people, and the extent of innate endowment. However, it is generally agreed that whatever innate constraints do exist are a product of evolution. That is, for each such constraint, natural selection must have preferred those organisms that were endowed with it. Thus, innate constraints become incorporated into an organism on a different timescale, and by a different process, than do constraints acquired by individuals via learning. Connectionist modeling provides a fruitful medium for exploring the relationship between evolution and learning. If strings of symbols are used to specify a variety of networks with different connection patterns and initial weights, a procedure known as the *genetic algorithm* (Holland, 1975) can operate on the strings so as to simulate natural selection. Since the strings specify the structure of networks, the success of the networks at learning can be used to specify a fitness function for the strings, and the strings can be permitted to undergo selective reproduction with random mutation. Nolfi, Elman, and Parisi (1994) not only showed that evolution and learning could be integrated in connectionist networks in this way, but provided a connectionist explanation of the puzzling Baldwin effect—the finding that successful learners will do better in evolutionary competition even though what is learned is not inherited (Baldwin, 1896). In subsequent research, Nolfi, Miglino, and Parisi (1994) added a developmental component in which the strings only specified general constraints for networks whose actual development also depended on the environment in which they developed. (The task presented to the networks was to control sensor-guided movement in tiny robots.) In simulating adaptive change at three different timescales—evolution, development, and learning—this project provided a nuanced exploration of the interaction between native endowment and learning.

This hardly scratches the surface of how connectionist modeling can be used to advance an interactionist perspective, but we must proceed to the constructivism that also characterizes the third contender and is even more crucial to its success. In Piaget's words: "The essential functions of the mind consist in understanding and in inventing, in other words, in *building up structures* by structuring reality." (Piaget, 1971, p. 27) In another context (an interview reported in Bringuier, 1977, p. 63), he emphasized how the construction process builds something that is initially not found either in the mind or in the external world: "I think that all *structures are constructed* and that the fundamental feature is the course of this construction: Nothing is given at the start, except some limiting points on which all the rest is based. The structures are neither given in advance in the human mind nor in the external world, as we perceive or organize it".

Piaget also proposed processes that not only moved the developing system towards increasingly complex structures but also continued to operate throughout life, in particular, assimilation, accommodation, and equilibration. He offered very detailed accounts of how these processes could be glimpsed in the ways that individual children grappled with the ingenious questions he posed to them. Piaget's severest critics focused, however, not on specific proposals of this sort but rather on the futility of the endeavor itself. Fodor (1980) voiced this impossibility argument: "It is *never* possible to learn a richer logic on the basis of a weaker logic, if what you mean by learning is hypothesis formation and confirmation." As in his 1975 argument, the foundation for Fodor's criticism of Piaget was the assumption that learning is a process of hypotheses testing, and that to test a hypothesis it must be possible to represent it. If you can represent it in the previously existing system, than the resulting system cannot represent anything that the previous system could not already represent.

Although Fodor's objection is restricted to approaches that construe learning as hypothesis testing, it serves an important function of making clear one of the essential features for an adequate constructivist alternative to nativism and empiricism: it must reject the characterization of learning as hypothesis testing and develop a suitable alternative that involves the construction of representations. Moreover, it must be a framework that does not view development as merely maturation of pre-existing structures, since then all the directions are assumed to be laid down in advance, which is all the nativist requires. Specifically, it must characterize the representational capacities of the cognitive system as changing with development so that what is learned is not restricted to what can be expressed in terms of the initial representations.

The metaphor of constructivism points to an important element in the challenge of establishing a third option. It emphasizes putting together components to build something that the components alone cannot do. It is worth noting that this is a general feature of mechanisms—mechanisms perform their tasks as a result of the coordinated operation of their components and a whole mechanism does something that the components individually cannot do. In mechanisms built by engineers, what a given engineer contributes is a new means of organizing components so as to accomplish a task that previous mechanisms could not perform. For discovering such organization a successful engineer wins patents and awards, monetary or otherwise. It is not expected that engineers begin with the ultimate atoms of the universe to build something new. Rather, they start with existing components and organize them in novel ways. When they are most successful, they produce a device capable of tasks which previously seemed impossible.

This comparison with engineers designing mechanisms is suggestive of what is needed to achieve something new, but one might reject its applicability to the nativism debate by emphasizing the role the engineer's cognitive system plays in the design process. The engineer already has a representational system capable of representing the new structure. That no one has formed that representation previously is simply an indication of the rich capacities of a representational system that permits construction of an infinite number of representations. To provide a plausible alternative to nativism, it must be shown that it is possible, even in the absence such a pre-existing representational system, to build a more powerful mechanism from a weaker one. More specifically, in the context of psychology, one must show that it is possible to build a more powerful representational system from a weaker one without relying on an external designer who already has such a representational system. (It is worth noting that a variation of this argument might be invoked to argue against the capacity of natural selection to produce novelty. The representational capacity of the genetic code is fixed and what occurs in evolution is merely the putting together and selection of new combinations of representations. Thus, evolution also cannot produce anything beyond what is already present. The nativist who raises Fodor's objection and who wants to appeal to evolution to explain innate endowments needs to be careful not to prove too much!)

Making Constructivism Work

For constructivism to meet the challenge set by nativism and constitute a viable third way that isn't simply a version of empiricism, a procedure must be identified for constructing something that is more powerful than the components out of which it was constructed. Recent work in chemistry and biology has been in the business of investigating just such systems—systems that exhibit the property of *self-organization*. Perhaps the best known of these systems is the B-Z reaction, a reaction discovered in 1958 by the Russian chemist Boris P. Belousov. Belousov had been investigating the Krebs Cycle (an important series of chemical reactions involved in metabolism) when he observed that a solution of citric acid, acidified bromate, and a ceric salt oscillated between appearing yellow and appearing clear. The idea of oscillatory phenomena violated the general conceptual framework in chemistry in which reactions proceed until a stable configuration is obtained, and Belousov's paper initially was rejected for publication on the grounds that what he reported was impossible. He was able to publish the paper only in a relatively obscure volume of proceedings from a conference. Subsequently, biophysicist Anatol M. Zhabotinsky developed a variant on Belousov's reaction, using malonic acid rather than citric acid, and showed that when a thin layer of reactants is left undisturbed, varying geometric patterns such as concentric circles and spirals propagate across the medium. Although initially merely a curiosity, the observed oscillations eventually were given an account in a detailed model of a mechanism for generating them proposed by Richard M. Noyes, Richard J. Field, and Endre Koros (Field, Koros, & Noyes, 1972). Importantly, this model showed how relatively simple components could organize themselves in such complex ways.

The geometric patterns produced in the B-Z reaction are transitory and are not the building blocks of anything greater than the reaction itself. Nonetheless, the reaction exemplifies concepts that have broader generality. The reason that the B-Z reaction was surprising was that we generally think in terms of linear processes which proceed to a stable terminal condition. But many systems in nature are non-linear and involve processes in which the product of the operation of one component feeds back onto an operation of another component conceptualized as operating earlier in the system. Such feedback can be of two sorts: negative feedback that depresses the activity of the earlier component or positive feedback that increases its activity. Once negative feedback processes were found in biological systems, researchers, especially those involved in the cybernetics movement, saw the significance of negative feedback in providing the capacity for a mechanism to regulate its performance, producing a product only when it was needed (Rosenblueth, Wiener, & Bigelow, 1943; Wiener, 1948). Positive feedback, on the other hand, seems to lead to run-away behavior and thus has been considered by many theorists to be something to be avoided. However, recently theorists investigating the origins of life have pointed to the possible importance of autocatalytic sets in living systems. In an autocatalytic set, one reaction produces a product which catalyzes another reaction, whose product in turn catalyzes the first reaction. In such a situation, if the proper materials occur together, the reaction will be able to sustain itself—assuming, of course, a source of energy sufficient to maintain the reactions (Kaufmann, 1993).

Self organizing auto-catalytic sets are not capable of maintaining themselves on their own. To achieve that, such a system also needs to be autopoietic—that is, capable of making and remaking itself. The notion of autopoiesis was developed by Maturana and Varela, who characterize it in the following way:

> *An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which:*
>
> > (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and

> (ii) constitute it (the machine) as a concrete entity in the space in which they (the components) exist by specifying the topological domain of its realization as such a network (Maturana & Varela, 1980, p. 78-79).

The details of autocatalysis and autopoiesis are not critical for present purposes. What is important is the departure from focusing only on linear organization and the recognition of non-linearity at the foundation of living systems. Such non-linear organization, involving both negative and positive feedback, creates the possibility of self-organizing complex systems that accomplish more than any of their components are capable of achieving. Moreover, such systems do not require a designer that already has the representational powers to represent them.

The notions of autocatalysis and autopoiesis were introduced in the context of biological systems, and the application of these concepts to the mental domain is clearly an extension. But they point the way to developing an account of constructivism that is neither nativist nor empiricist. The key is to construe the development of mental representations as involving the integration of components into novel, self-sustaining structures. At this stage we do not know enough about higher mental processes to establish in detail how representations are constructed through self-organizing processes. But two examples are suggestive of how self-organizing systems can achieve representational capacities beyond those comprised of linear strings of pre-existing representations. (For other suggestions for developing a viable constructivist alternative to the nativism-empiricism dichotomy, see Quartz & Sejnowski, 1997; Quartz, 1999.)

The first example focuses on the fact that in viewing ambiguous figures such as the Necker cube, subjects spontaneously and repeatedly shift from one interpretation to another. Using coupled map lattices whose internal dynamics are chaotic, van Leeuwen, Steyvers, and Nooter demonstrated that self-organization could lead to a metastable synchronization of oscillatory units which constituted one interpretation of the ambiguous figure. Such a state was only metastable in that the chaotic dynamics of the individual units would lead the synchrony to break down spontaneously. This would enable the network to explore more of its state space and then settle into another metastable synchronization that constituted another interpretation of the figure. In this case, self-organization among components interacting non-linearly permitted a system to develop representations of relations between elements of a visual pattern and to shift between alternative representations without such representations being built into the system and without specifying rules for shifting between representations.

A quite different example of self-organizing representations is found in Jeffrey Elman's work with recurrent neural networks learning simple grammars. Recurrent networks are connectionist networks in which the activities of hidden units (or in some versions, output units) are recycled as part of the input when a second item (e.g., a second word in a sentence) is presented. Such recurrent processing enables the network to bias its response to a given input in light of recent inputs. Elman trained his networks to predict the next word in a corpus of text constructed from a grammar that permitted relatively complex forms such as multiple relative clauses, each with its own verb phrases (Elman, 1991). An interesting result Elman obtained was that when the network was simply trained on a corpus in which simple and complex sentences were mixed, it failed to learn. But when it was first trained on simple sentences and then more complex sentences were added to its training corpus, it reached a high level of performance. Elman refers to this as *starting small* and he argues that a to learn a grammar that is complex enough to include multiply embedded clauses, a system is aided by first developing representations of simple sentences and then adapting these to represent more complex sentences. With this kind of training regimen, his networks could bind each verb to the appropriate subject noun phrase (Elman, 1993). (The technique of restricting the network to processing simple sentences is very different from the way in which human children learn language, but Elman shows that the same effects are obtained if instead the recurrent activations were initially limited. This may roughly correspond to the limits on working memory in the developing brains of young children. Elman's argument is that the capacity to learn complex systems such as grammars for natural languages may depend upon the opportunity to learn under restricted (simplifying) conditions before confronting the full range of complexity in a corpus.

Elman's success depended upon a reorganization of existing representations achieved in the course of non-linear dynamical processing within a connectionist network, not just the logical structuring of existing representations (as might be realized in a traditional artificial intelligence system). This permitted the basic knowledge of which (implicit) classes of nouns are bound to which (implicit) classes of verbs to carry over and guide performance with more complex grammatical structures. Using techniques such as principal components analysis, Elman was able to show how the network produced subtly different representations of the same clause depending on how deeply it was embedded. Thus, the network was able to handle multiply-embedded sentences, for example, *boy chases boy who chases boy who chases boy*. To achieve this, Elman designed fairly complex networks (there were three layers of hidden units, the largest of which had 70 units). Although this makes detailed analysis of the dynamics of the network challenging, the inclusion of hidden layers and especially of recurrent connections made this a persuasive example of the emergence of representations from self-organizing processes.

These context-sensitive representations of words enabled the network to process complex linguistic structures of the type that Miller and Chomsky (1963) had argued could not be handled by statistical inference procedures alone. This further suggests that Elman's networks are using representations that are not just simple composites of atomic representations, but ones that enable the sorts of processing for which structured rules are often taken to be necessary. Even more impressive performances were obtained when Morten Christiansen built on Elman's results. Christiansen trained networks on more complex grammars that permitted such structures as multiply center-embedded sentences and demonstrated that the networks began to make errors at approximately the same number of embeddings as human subject do. The self-organizing processes in these networks seem to have resulted in the *construction* of representational capacities from weaker ones, the sort of construction Fodor denied was possible.

Conclusion

One reason development, both biological and psychological, is relegated to a secondary status is that it seems quite reasonable to explain the mechanism responsible for a phenomenon without investigating its development. Moreover, if one does ask about the origins of the mechanism, one seems forced to accept either the nativist view that the mechanism is innate or the empiricist view that it was created from simple associations of simpler components. Piaget's third alternative—an interactionist/constructivist approach—has been regarded by many nativists as no better than the empiricist alternative and to face the hopeless task of showing how a structure can be composed that has greater powers than its components. I have argued that the way to surmount this problem is to emphasize the role organization plays in a mechanism—it is the organization that enables a mechanism to accomplish more than its parts can. But to make this work one must offer an account of how such organization can arise without being previously represented (e.g., in a cognitive representation or in a genetic system). I have sketched how self-organizing systems that take advantage of nonlinear positive and negative feedback provide the tools for such construction of a mechanism that can accomplish more than its parts are able, and offered two examples from recent cognitive modeling of how such self-organization can enable a system to develop representations not already present in the system.

References

Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist, 30*, 441-451.

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford.

Bechtel, W. (2003). Modules, brain parts, and evolutionary psychology. In F. Rauscher (Ed.), *Evolutionary psychology: Alternative approaches*. Dordrecht: Kluwer.

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

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

Bringuier, J. C. (1977). *Conversations libres avec Jean Piaget*. Paris: R. Laffont.

Chomsky, N. (1959). Review of *Verbal Behavior*. *Language, 35*, 26-58.

Chomsky, N. (1980). Rules and representations. *The Behavioral and Brain Sciences, 3*, 1-61.

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.

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

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Field, R. J., Koros, E., & Noyes, R. M. (1972). Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the Bromate–Cerium–Malonic acid system. *Journal of the American Chemical Society, 94*, 8649–8664.

Fodor, J. A. (1975). *The language of thought*. New York: Crowell.

Fodor, J. A. (1980). Fixation of belief and concept acquisition. In M. Piatelli-Palmarini (Ed.), *Language and learning: The debate between Jean Piaget and Noam Chomskv*. Cambridge, MA: Harvard University Press.

Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*, 183-204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.

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

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447-474.

Harvey, W. (1651). *Exercitationes*. London: Octavian Pulleyn.

Hempel, C. G. (1966). *Philosophy of natural science.* Englewood Cliffs, NJ:: Prentice-Hall.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*, 137-175.

Holland, J. H. (1975). *Adaptation and natural and artificial systems*. Cambridge, MA: MIT Press.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science, 15*, 219-250.

Kaufmann, S. A. (1993). *The origins of order*. Oxford: Oxford University Press.

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 F. J. Varela (Ed.), *Autopoiesis and Cognition: The Realization of the Living* (pp. 59-138). Dordrecht: D. Reidel.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce & R. R. Bush & E. Galantner (Eds.), *Handbook of mathematical psychology* (Vol. 2). New York: Wiley.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Nolfi, S., Elman, J. L., & Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behavior, 3*(1), 5-28.

Nolfi, S., Miglino, O., & Parisi, D. (1994). Phenotypic plasticity in evolving neural networks. In D. B. Gaussier & J.-D. Nocoud (Eds.), *Proceeding of the International Conference From Perception to Action* (pp. 146-157). Los Alamitos, CA: IEEE Press.

Piaget, J. (1971). *The science of education and the psychology of the child*. London: Longmans.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron. *Cognition, 38*, 1-60.

Quartz, S. R. (1999). The constructivist brain. *Trends in Cognitive Sciences, 3*, 48-57.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences, 20*, 537-596.

Rosenblatt, F. (1962). *Principles of neurodynamics*. New York: Sparan.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose, and teleology. *Philosophy of Science, 10*, 18-24.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2. Psychological and biological models*. 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.

Wiener, N. (1948). *Cybernetics: Or, control and communication in the animal machine*. New York: Wiley.

Wolff, C. F. (1759). *Theoria Generationis*: Halle and der Saale.

Wright, C., & Bechtel, W. (in press). Mechanisms. In P. Thagard (Ed.), *Philosophy of psychology and cognitive science*: Elsevier.