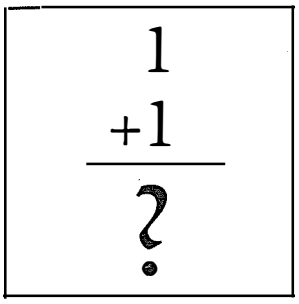

Cognitive Representation in the Brain

MARK H. BICKHARD, *University of Texas at Austin*

Glossary

Connectionism Approach to representation as being distributed in the patterns of activations of nodes within a network. Inspired in part by the high interconnectivity of the nervous system

Encodingism View that representation is fundamentally constituted as elements or structures that are in known correspondences with what they represent

Ensemble Population of active elements in which the statistical properties of single elements over time is equal to the statistical properties of the population of elements at a single time

Incoherence problem Impossibility of specifying what an encoding is supposed to represent except in terms of some other already available representation, and the incoherence that results when that regress is supposed to halt with foundational encodings

Information processing View of cognition as consisting of the processing—the manipulation, combination, and generation—of symbolic encodings

Interactivism Functional and emergent approach to representation. Representation emerges as interactive differentiations of environments and consequent indications of possible further system activity in the service of goal-directed interactions

Skepticism Argument that it is impossible to check the accuracy of our representations of the world because it is impossible to know anything about the world except in terms of those representations themselves. Any purported check, then, is checking the representations against themselves—it is circular.

Transduction Transformation of form of energy. Also, a supposed generation of sensory encodings from encounters with environmental energy

STUDIES OF THE COGNITIVE ASPECTS of brain functioning cannot proceed without assumptions concerning the nature of representation. Since the demise of classical associationism, those assumptions, both in cognitive neuroscience and in cognitive psychology in general, have been dominated by the computer-inspired information processing model. In this model, representations are taken as being constituted as symbolic encodings, which are generated, processed, and transmitted by the central nervous system. This model has dominated so long and so thoroughly that it at times has seemed to attain the level of unquestionable common sense—the way things obviously must be.

More recently, however, several competing alternatives to this standard position, and criticisms of these standard assumptions, have arisen. As a result, future studies will increasingly be forced to take into explicit account their conceptual assumptions concerning representation as well as the neurophysiological and psychological results against which their models are tested.

One important alternative to information processing views is that of connectionism or parallel distributed processing (PDP). Another position has recently emerged in robotics but has not yet had much impact in brain studies. This approach attempts to eschew representation altogether in favor of sys-

Encyclopedia of Human Biology, Volume 2.

tems without data structures that can nevertheless accomplish their goals.

All three of these positions, however—information processing and its two alternatives—*share* one basic assumption concerning the nature of representation, in spite of their differences in how that assumption is developed. The assumption in common among them is that representation is constituted as encodings, whether or not these are taken to be symbolic. That assumption is itself subject to severe criticism, which, in turn, leads to a fourth alternative to all three positions: an interactive conception of representation.

This article is primarily a review of the four positions concerning the nature of representation and some of the arguments among them. With respect to the information processing position especially, this will also involve some illustrative examples of how that approach can be applied to brain functioning. I will be arguing in favor of the interactivist position.

I. Information Processing

The backbone of the information processing perspective is the presumed flow of information from environment to perception to cognition to language. Information originates in the environment and is processed through the senses into the brain or mind, where further cognitive processing occurs and where new encodings of resultant mental contents can be generated and transmitted as language utterances. Those utterances, in turn, will be received by an audience and decoded in accordance with their semantics into cognitive contents for that audience. The basic flow of information, then, is from the environment, through the senses and cognition, and into the environment again as language—from which it will in general reenter the nervous system via the perception and understanding of language.

The various steps of this sequence are of three general kinds: the transduction of new encoding elements in the primary perceptual organs in response to encounters with environmental information, the generation of new encodings on the basis of already present encodings in the various processing steps (this is a form of heuristic and perhaps implicit inference of new encodings on the "premise" of extant encodings), and the emission of encodings via language. All three of these processes are being investigated, and knowledge of differentiations and specializations within the central nervous system by sensory modality and form of cognition is growing rapidly.

The sensory nervous system is generally considered to provide only two possible forms of basic encoding. The transduction process must result in signals being transmitted along various *axons* carrying some *frequency* of impulse. Basic sensory encodings, then, must be implemented in some combination of line (axon or spatial) and frequency (temporal) encoding.

For example, human color vision is based on three different types of receptors, each attuned to transduce a differing range of electromagnetic wavelengths. The transduction sensitivities of these types of receptors are maximal in, respectively, blue, green, and red ranges of color. This gives rise to a *line encoding* of *color,* which undergoes several stages of further processing in the retina, the visual pathways, and the visual cortex. Both color receptors, cones, and more general light intensity receptors, rods, are distributed spatically over the retina (with cones concentrated in the fovea), and this gives rise to a *line encoding* of relative *spatial position* of light reception. *Light intensity* itself receives a *frequency encoding*. The topology of these spatial relationships tends to be maintained through the several layers of further processing; thus, the spatial encoding of relative position of reception tends to be maintained. The auditory system for another example, yields primarily a *line encoding* of *frequency,* although lower frequencies seem to involve some degree of *frequency encoding*. [*See* COLOR VISION; EARS AND HEARING.]

These relatively simple correspondences between properties of the stimulus, on the one hand, and lines or frequencies of neural activity, on the other, become more complex and less well understood with progressive steps of processing. Some of the more complicated and well-known examples are the apparent motion and "feature" detectors of the visual system. The feature detectors seem to be sensitive to such features as edges and orientations—important properties of visual boundaries. [*See* VISUAL SYSTEM.]

Such models of perceptual encoding are based on single neurons and their activities as the elements of the presumed encodings. However much insight these models may provide for perceptual representation, there are strong reasons to think that single neurons are not the locus of representation in the central nervous system. One consideration is sim-

ply that the limiting case of such single neuron representation yields single neuron "encodings" of all of our concepts and representations—a common *reductio ad absurdum* of this is the infamous "grandmother neuron" that represents our grandmother. Since there is a continual loss of neurons to cell death, we would experience random and total losses of representation of whatever those neurons represented, including, potentially, our grandmothers. Such unitized representational losses are not found with neural death; thus single neurons cannot be the microanatomical locus of concept representation. Furthermore, activities of single cortical neurons in response to repeated stimulus instances are found in general to be highly unreliable, also not giving a foundation for a single neuron locus of representational encoding.

One design solution to this unreliability problem (cell death is itself a version of unreliability) is redundancy: if many neurons are serving the same representational function redundantly, then the loss of one or more, or the unreliability of all of them, can in principle be compensated for by the activity of the whole redundant set. A more powerful hypothesis, however, is that the functional unit is not the single neuron at all but, rather, local populations of neurons that function as statistical ensembles. The relevant properties of such ensembles would be the temporally and perhaps spatially organized patterns of oscillations within the ensembles, which would modulate the similar activities of other ensembles. Cognitive activity would consist of such modulatory processes disseminating and interacting throughout the brain.

This is a very different notion from the classical "switchboard" model of brain activity in which impulses are generated at perceptual surfaces and then switched to various output neurons within the central nervous system (CNS). In this long outdated switchboard model, even the notion of frequency encoding is distorted in that the switchboard metaphor emphasizes on and off relationships, not frequencies. The ensemble model builds on the spatial and frequency characteristics of neural functioning, provides a redundancy with respect to single level neurons in that the ensembles will exhibit reliable statistical properties of their oscillations in spite of single neuron unreliability (this is in effect a reduction of noise or error variance via a larger sample), and, for the first time among the models discussed, acknowledges the endogenous activity of the central nervous system.

This later point is potentially quite important. Neurons in general are not quiescent until stimulated by synaptic transmissions or sensory input. Neurons exhibit baseline frequencies of axonal impulses, varying from neural type to neural type, and varying from zero to high frequency. This intrinsic neural activity is continuous. Sensory inputs do not switch this on or off so much as they modulate the frequencies and patterns in which this ongoing activity takes place. The ensemble notion is a population level version of this basic point concerning even single neurons.

Perception as a modulation of internal endogenous activity is a quite different notion from that suggested by the simple information processing flow from environment to perception to cognition. Modulation is not the same relationship as simple encoded input. Even the information processing models, however, acknowledge the necessity of contributions to perception and cognition from previously learned or innate sources—the sensory information is not adequate to cognition, nor, most models hold, even to perception. Memories, for example, might be postulated as involved in the inferences from basic sensory reception to full perceptions of objects located and moving in space and time. Modulation of ongoing activity, then, is not at such deep variance with such versions of the information processing approach. [*See* PERCEPTION.]

Even these forms of information processing models, however, retain the presupposition that all cognition is ultimately input from the environment (or perhaps provided innately). If not present in current sensory input, relevant information must have been provided in earlier experiences and be available in memory. That is, such models, except for the "out" of innatism, force an empiricist epistemology, in which the senses are the source of all knowledge. Empiricist epistemologies have not fared well in epistemology or the philosophy of science. Understanding the necessity of mathematical relationships, such as $1 + 1 = 2$, for example, has been a classical counter to empiricisms—it might be conceivable that we could learn *that* $1 + 1 = 2$ simply from experience, but no amount of experience will ever provide knowledge that this relationship is logically necessary. Mathematics would remain on a par with, say, astronomy, in which the number of planets also remains consistent, no matter how many times you look, but that number is *not* necessary. For that matter, there does not seem to be any perceptual realm at all for mathematics—we

can see pebbles but not numbers. Many other cognitions, such as of virtues and vices, are not presentable in sensory form. Such considerations suggest that the endogenous activity of the CNS is not simply bringing to bear previously perceived information but that it is involved in some way in *emergent* representational phenomena.

A convergent consideration for the notion of endogenously active ensembles as units of functional activity is the acknowledgment that the organism is fundamentally engaged in physical activity in the environment. Such interaction with the world requires in most cases *correct timing* of the organism's side of the interactions. By correct timing is meant neither too fast nor too slow nor in the wrong phase. Walking, for example, is not so much a matter of pushing the legs back and forth as it is a matter of exciting and modulating an intrinsic oscillation in the spinal cord and of the skeletal–musculatory system itself. Most activities require such timing—driving a car, catching a ball, running, and so on—and timing is fundamentally based on oscillatory phenomena. We would expect, therefore, that oscillations and modulations would be fundamental to the operating design principles of the nervous system. Ensembles, then, not only provide an answer to the problem of unreliability and pose the problem and the promise of intrinsic endogenous activity, they are also endowed with the basic solution to the problem of timing in action and interaction.

The information processing approach has no intrinsic place for timing. The sequence of operations on symbolic encodings has all the same *formal* properties no matter what the timing may be of the steps involved in that sequence. It is clear that results may be obtained too late to be of any good, and, thus, that *speed* is necessary, but in this view, timing is irrelevant to the nature of cognitive activity per se. Cognition abstracts away from the timing considerations that are essential to action, according to this view, but even so we could expect the timing properties of oscillatory phenomena to dominate the functioning of the CNS. I will argue later that timing is in fact *not* irrelevant to cognition in general.

II. Connectionism

Considerations of the highly parallel manner of functioning of neurons and neural circuitry, and of the enormously complex interconnectedness of neural circuitry, contributed to the inspiration of one major alternative to standard information processing approaches—connectionism or parallel distributed processing. The underlying metaphor for the information processing approach is the von Neumann computer, which has only one locus of processing. Parallel processing in computers can be introduced by multiplying the number of simultaneously active processing units, but something more than this seemed to be taking place in the brain.

One highly persuasive consideration is that the brain accomplishes many tasks, such as various forms of pattern recognition, in a short amount of time and, therefore, in a small number of strictly sequential "steps" of processing. One potential solution was to posit that the many steps that seemed to be logically required were carried out simultaneously and in parallel in multiple processing units. This solution, however, retained the basic assumptions of the information processing approach and simply introduced multiple processors, and it was a conceivable approach only when the basic task did not involve internal dependencies that required sequential processing steps—only when the processing could in fact be broken down into multiple parallel streams.

Another strong consideration was that standard information processing approaches had failed miserably at modeling the phenomena of learning. Systems could be designed that succeeded in "learning" things that were close to what they had been designed to solve, but any significant generalization beyond the problem for which they were designed seemed unattainable. One perspective on this failure comes from noting that the information processing approach construes all processing in terms of the generation and elimination of *instances* of various *types* of encoding elements, but the basic types of encodings themselves must be designed in from the beginning—there is no way to generate new types of symbolic encoding representations.

The major excitement of the connectionist or PDP approaches is that they seem to solve this problem of learning, of the generation of new representations. PDP systems can undergo training with respect to sets of problems and generate solutions to them that then generalize beyond the training set. At times, the form of the generalizations found seems tantalizingly similar to human solutions to those same problems.

A PDP system engages in two levels of dynamic activity. The primary level is that activity by which a categorization, the representation, of an input array is settled upon. The secondary or meta-level of activity is that by which the system "learns" to correctly categorize such input patterns.

A PDP system is a set of functional nodes, each of which is capable of varying levels of "activation," interconnected by a fixed topology of paths. Each connection between nodes has a weight, which can be positive or negative. The nodes and interconnections are frequently organized into layers, perhaps with feedback among layers. Some subset of nodes are connected directly to the environment, from which they receive "activation," and they in turn activate the nodes to which they are connected in accordance with the weights of the connection paths. These nodes activate still further nodes in accordance with their connectivities and weights, and so on. The system eventually settles down into a fixed pattern of levels of activation in the nodes, or in some selected subset of the nodes, that is specific to that particular pattern of inputs. The first key to the power and appeal of PDP models is that that pattern of resultant activations can be taken as a classification of the input pattern: it will classify *together* all such input patterns that yield that same resultant activation pattern and will classify as *different* all input patterns that result in some different final activation pattern. The class of possible final activation patterns, then, forms the class of classification *categories* for the possible input patterns. Note that the processes of settling of the node activations is massively parallel among all the nodes and their weighted interconnectivities.

The second, and most important, key to the appeal of PDP models is that various adjustment rules can be used to adjust the weights of the connections among the nodes, in accordance with various training "experiences," in order to "learn" input pattern classifications. The organization of connections remains fixed in such training, but the changes in the *weights* of those connections can change the entire first level—classifying—dynamics of the overall system. In particular, it can result in differing resultant classifications of the input patterns. The system, in other words, can adjust to, can be trained to, new and desired classification schemes. Insofar as the input pattern-classifying activation patterns are taken to be *representations* of those categories of input patterns, the system can be construed as generating new representations of novel input categories, something that is impossible in the standard information processing approach.

With proper design and appropriate shifts in interpretation, PDP systems can manifest still other characteristics that are simultaneously powerful, exciting, and reminiscent of the way in which the brain functions. One important example derives from the possibility of the *input* activation pattern being any of several *sub*patterns of the overall *resultant* activation pattern—so that any piece of the overall pattern as input results in the activation of the whole pattern—in which case we have a model of content addressable memory. Content addressable memory is a form of memory that permits memory representations to be accessed directly in terms of their representational *contents,* rather than just in terms of their *location* in the memory organization. Human memory, in particular, manifests this phenomenon.

A different shift in interpretation considers the input patterns themselves to be whole patterns, but the resultant activation pattern to be a composite of the permitted input patterns. Under this interpretation, the system manifests an *association* between the various input patterns—an associative memory, again manifested in human memory.

Connectionist approaches, then, capture a parallelism at least reminiscent of the functioning of the brain, model the emergence of new categorizations, model a form of content-addressable memory and associative memory, and other properties of human memory. They are an exciting alternative to information processing approaches for these and additional reasons and are being pursued eagerly.

They are not without their critics, however. One of the most powerful criticisms of the potentialities of PDP approaches turns on what from another perspective is one of their greatest strengths—the singularity and lack of internal structure of the categorizing patterns of activation. This is an aspect of their greatest strength in that emergent novel representations would be expected to be singular and without internal representational structure. To simply put together already available representations in some new structure is what information processing approaches already do and does not constitute emergent novel representation. On the other hand, it is precisely the ability to construct new *structures* of already available representations and, thus, to implicitly capture not only the resultant representation but also its internal representational structure that is the *forte* of symbolic encoding information

processing approaches. It is argued that for both language and cognition alike, this componentiality of representation is necessary and is not provided by PDP approaches. For example, any genuine cognitive system, so the argument goes, that is capable of thinking "John loves Mary" is also capable of thinking "Mary loves John." This generalization of ability is natural in a symbolic encoding framework but not in the more holistic PDP framework. Needless to say, the arguments and explorations continue.

One obvious notion, for example, is the possibility of hybrid systems in which a basic PDP-type layer provides the representational categories that can then be operated on and processed in a more conventional information processing manner. Whether or not such is feasible, and what might be gained, remains to be explored.

Connectionism boasts a natural manifestation of several inherent properties of CNS functioning: parallelism, emergentism, content addressability and associativity, and so on. Nevertheless, there are a number of inadequacies, or at lest disanalogies, of connectionist approaches with respect to this comparison. For example, PDP networks "represent" by virtue of static patterns of activation of the nodes, once settled into, while the CNS is engaged in continuous ongoing activity. It is at least plausible, and even likely, that cognition in the brain is a function of that activity and cannot be captured in such static models. In this respect, among others, PDP networks are *not* akin to neural ensembles. A similar observation is that the CNS is engaged in interaction with an environment (internal or external), while a PDP network has no comparable outputs at all. Furthermore, although PDP networks do manifest a kind of emergence of categorization abilities, the "learning" rules by which this is accomplished are relatively inefficient and are, in general, *not* plausible as models of learning in the brain.

III. Robotics

The information processing approach encounters many problems of interpretation. One important example is what is known as the empty symbol problem. The basic notion underlying this problem involves the sense in which encoded symbols in the information processing approach are supposed to represent various events and objects and facts in the environment by virtue of being in correspondence with those events and objects and facts. Transduction, for example, is fundamentally a change in form of energy from some environmental form to some form of neural activity. There is nothing epistemic or representational in this energy-change level of consideration—such changes in form of energy or activity occur ubiquitously in the physical world, without being confused with representation. Transduction in sensory systems, however, is taken not only to be constituted by such energy changes and their resulting correspondences, but those correspondences, in turn, are taken as representing whatever those correspondences are with, whatever was in fact transduced. The empty symbol problem arises, among other ways, from consideration of how those factual correspondences could represent what has been transduced or some other relevant aspect of the correspondence. Specifically, how does the system know what the correspondences are with? Or, among the multitude of things that are in fact in correspondence—light patterns, quantum electron processes in the surfaces of objects, chemical reactions in the retina, and so on—how does the system know which are being represented? The scientist–observer can analyze these correspondences and analyze as well which of those correspondences seem to be ecologically relevant, but this only establishes which correspondences do occur and which of those would be ecologically desirable to represent. There does not seem to be any way for the system itself to have epistemic contact with whatever it is in causal contact with, to obtain transduced *encodings,* not just transduced *energies.* The internal symbols, in other words, seem doomed to be empty. They differ in shape or size or some other formal properties that allow them to be differentiated and operated upon, but they carry no representational content, or, at least, it is not understood how they can carry any representational content *for the system itself.*

Because of this and related difficulties in the information processing approach, some researchers in robotics have proposed that representation be avoided altogether and, furthermore, that robots can function quite well without any representation at all. Insofar as that is correct, it may be that symbols and data structures are superfluous for robotics, that representation is the wrong level or the wrong sort of abstraction for robots.

In place of notions of representation communica-

tion and processing, design is in terms of minimally coordinated activity systems, each of which is competent to its own perceptions and actions. An energy transducer, for example, doesn't have to establish an encoding of a wall, as long as it controls the locomotion of the system to avoid bumping into walls.

In important ways, this is a return to the roots of cybernetics and control theory out of which computer information models evolved. The control of effective interaction with an environment has largely been lost from information processing approaches as being of any interest beyond that of robotic engineering. In particular, it is not generally understood to have any basic relevance to foundational issues of representation or cognitive science in general. Roboticists, clearly, have not been able to ignore such concerns quite so readily.

This antirepresentationalism of some roboticists emphasizes the interactive and hierarchically organized control structure aspects of the nervous system—aspects which are absent in both the information processing and the PDP approaches. I will be arguing that these aspects are not only of practical design relevance but that they are intrinsic to the nature of cognition and representation as well.

IV. The Encodingism Commonality

There is a common notion of representation among the three approaches. It is that representation is constituted as encodings of what is to be represented. In the information processing approach, this is a direct assumption, with the basic atomic encoding types designed directly into the system. In the PDP approach, these basic encodings are presumed to emerge in the learning process of the PDP network, but what is learned is still a correspondence between patterns that is presumed to constitute a representation—an encoding. The antirepresentationalism position accepts the same notions of encoding representation but concludes, not that they are wrong, but that robotics can proceed without them. I will argue that all three approaches are in error in this common assumption.

V. What's Wrong with Encodingism?

A prototypic encoding is a representational stand-in. It is some element that is specialized to serve a representational function, to carry a representational content, that is determined by some *other* representation—which may also be an encoding—thus, it "stands-in" for that other representation. In Morse code, for example, ". . ." stands-in for "S," while bit patterns serve the stand-in function in computers. In this sense, encodings most certainly exist and are quite powerful and useful. They specialize and differentiate representational functions and change the form of representational elements in such ways as to allow processing to occur that would otherwise be difficult or impossible: ". . ." can be sent over a telegraph wire, while "S" cannot.

The term "encoding," however, is also used in a variety of derivative ways that do not comport with these paradigmatic cases. Genes, for example, are often called encodings of the proteins whose construction they control, but they are not *representations* in any legitimate sense: instead, DNA base pair triples and strings of such are elements of a complex control organization that builds proteins. The selectivity of those DNA triples for certain amino acids in the control process is what motivates their being deemed encodings—there is a correspondence involved. This generalization of the paradigmatic encoding notion makes quite clear the seductive power of the correspondence notion, even though, in this case, there is no epistemic or knowing or representing agent at all. It is only the stand-in notion of encoding that is a representation, and, therefore, it is only this notion that I will analyze in terms of its sufficiency for the general notion of representation.

The existence and importance of encodings is not at issue. What is at issue is the assumption that representation is fully characterized by encodings. There is a complex of related criticisms and arguments against strictly encodingist conceptions of representations, some of which are of ancient provenance. Several of the core components of this complex of arguments will be summarized. I submit that these arguments render any simple encodingism logically incoherent: strict encodingisms cannot make logical sense, and certainly cannot ground models of cognition, at the neural level or more abstractly. That is, strict encodingism is not merely factually false, it logically cannot be true.

The first argument is that of skepticism. The basic skeptical argument first notes that in order to check whether or not our representations are correct we would have to compare those encodings against the

world that they are taken to represent. But since we can know the world which is supposedly being represented only via those representations themselves, we cannot ever get independent epistemic access to the world to check our encoded representations of it. The conclusion, then, is the classical skeptical contention that we cannot have genuine knowledge of the world since we cannot ever check the accuracy of our representations of it.

The second argument asks not about the accuracy of our representations but about the construction of them in the first place. The point is that in order to construct elements that are in correspondence with the world, we must already know what those correspondences are to be with, but that cannot occur until the correspondent encodings are constructed. Therefore, there is no way to get started; no way to know what encodings to construct. We must already represent the world before we can construct our representations of it.

A closely related consideration is to note that the *factual* correspondences found in sensory transduction between neural activity and environmental events do *not* establish *epistemic* correspondences. The problem concerning how to know which encodings to construct leads in this context to the question of how the system, the CNS, could possibly know what those sensory correspondences are with and, therefore, what those sensory elements are taken to be encodings of. In other words, how can the system turn nonrepresentational energy transductions into representational encodings? It would have to know what the internal neural activity was in correspondence with in the world in order to know what the representational content of that neural activity should be taken to be. It would have to already have its representation of the world in order to construct its representation of the world.

A quick apparent answer to these questions is to mention evolution and assume that they have been solved there. But the problem is logical, and evolution has no more power to escape them than does maturation, learning, or development. Note that the power of evolution to construct active systems that successfully interact with their environments, along the lines of the antirepresentational robotics position, is *not* questioned by these arguments. What *is* put into question is the ability to construe those transductions as more than useful control signals, to construe them as encodings. Where does their representational power, their representational content, come from, and how does it come into being? The factual correspondences that obtain between the environment and neural activity help explain how and why that neural activity is in fact useful, but that useful functioning does not require any encoded representations. How does, or could, evolution, maturation, learning, development, human design, or any other constructive process, construct representations out of control organizations, or out of anything else, that is not representational already?

An additional level of consideration derives not from the question of the accuracy of encoded representations, nor from the question of which ones to construct, but from the question of how the system could possibly know what any encodings in a strict encodingist system were even supposed to represent, before such questions of accuracy or rational construction. For actual encodings, the answer to this question is provided by the representation for which the encoding is a stand-in: the encoding represents the same thing as that for which it stands in. But this introduces a regress into the origin of the representational contents involved: they might be provided for encoding X in terms of encoding Y, and for Y in terms of Z, and so on, but this regress must end in a finite number of steps. If we consider a supposed grounding level of encodings, that are not stand-ins for any other representations, that are logically independent encoding representations, we find that there is no way to provide the necessary representational content. For some purported grounding level encoding "X," we might attempt to define it in terms of some other representations, in which case it would not be at the ground level, contrary to hypothesis. But this leaves us at best with "'X' represents whatever it is that 'X' represents," and this does not establish "X" as a representation of anything. The requirement for logically independent encodings in order to ground any strict encodingist model encounters a logical incoherence: logically independent encodings cannot exist.

The underlying reason for all of these problems with encodingisms is that the notion of encodings focuses on, and is enormously powerful for, change of form of representations *that already exist*, and for combinations of representations *that already exist*. There is nothing in the notion of encodings that can explain the *origin* of representation, the *emergence* of representation out of a ground that is itself not already representational. This is so whether that emergence is evolutionary, maturational, learning, developmental, or by design. Encodingisms model things that can be done with representations that

are already available, so to assume that representation in the broad sense can be fully captured in a strictly encodingist model is intrinsically incoherent. It encounters a requirement that encodings cannot serve—the requirement to explain representational emergence.

The closest attempt at such explanation within the encodingist framework is that of transduction, or its closely related notion of induction—a temporally extended transduction. But, as we have seen, these do not work. In both cases, the knower must already have the representation—the cognitive category or sensory encoding element—before it can notice or detect or transduce or postulate that "corresponded-with" element of category for its environment. Something more is needed, something that can account for representational emergence.

VI. Interactivism

Consider a system, or subsystem, in interaction with its environment. The internal course of that interaction will, in general, depend both on the internal organization of the system and on the particulars of the environment being interacted with. At the end of the interaction, the system will be left in some internal state, say state *A* or state *B*. If *A* and *B* are the only two possible final states of this system or subsystem, then those internal states will serve a function of *differentiating* possible environments into two classes—those environments that leave the system in *A*, and those that leave the system in *B*. A simple version of this differentiation involves systems or subsystems that have no outputs; they passively arrive at final internal states, e.g., energy transductions.

Note that at this point, the typical encodingist move would be to note that such differentiation establishes correspondences with the differentiated environments, and, therefore, *A* and *B encode* their respective correspondent categories of environments. The first step in this move is correct: the differentiations do establish factual correspondences with whatever it is that is differentiated. The second step is invalid: those factual correspondences do not in themselves constitute encoded representations of what the correspondences are with, of what the differentiations are differentiations of. The differentiations are more primitive than encodings, yet they do involve factual correspondences. It should be clear that the correspondences noted in actual sensory systems are in fact useful as differentiations and are epistemically nothing more than differentiations—all the system has functional access to is that state *A* is different from state *B* and that it is currently in state *A*. They are not encodings.

At this point, we are roughly in the position of the antirepresentationalist roboticists—potentially useful signals in a control structure—but with the recognition that something more is necessary. In particular, the antirepresentationalism of the roboticist position *accepts* the basic encodingist notions of representation, and there is independent reason to conclude that those notions are false and incoherent. The emergence of representation must itself be accounted for, and encodingism cannot do that.

We already have the emergence of a representational sort of function, differentiation, out of system organization that is not itself representational. What is yet missing is the emergence of representational content. Purely differentiating states *A* and *B* are truly empty; they differentiate, but, in standard senses, they represent nothing. The next task is to account for the emergence of their having representational content.

Suppose that the system with final states *A* and *B* is a subsystem of a larger goal-directed system. In this larger system, in general, various alternative strategies and heuristics will be available for attempting to reach various possible goals. In such a case, given some particular goal at a particular time, some selection among possible strategies and heuristics must be made. It may be that the system makes a choice of strategies or heuristics in part on the basis of whether the differentiating subsystem has reached final state *A* or *B*. If so—if, when attempting to reach goal *G72* and final state *A* obtains, try strategy *S17,* while if final state *B* obtains, try strategy *S46*—then such functional connections constitute implicit predications concerning the environments differentiated by *A* and *B*. In particular, the predications involved are "state *A* environments are appropriate to strategy *S17*; state *B* environments are appropriate to strategy *S46*." Such implicit predications are *about* the environments and can be true or false about those environments. They constitute *representations* of supposed environmental properties. They constitute *representational contents* attached to final states *A* and *B*.

Most importantly, they constitute *emergent* representational contents: there is nothing in the sys-

tem organizations involved that is itself a representation nor that is representational in a more general functional sense. The function of representation emerges in the further selection of system activity on the basis of initial environmental differentiations.

These considerations suffice to show: (1) that the interactive model of representation suffices to account for at least one form, a functional and implicit form, of representation, (2) that this form is capable of emergence from nonrepresentational ground, and (3) that this form is not itself an encoding form. There remain, of course, many questions concerning the interactive approach to representation. Among the most important are those concerned with the adequacy of interactive representation to the many representational phenomena. One important version of the adequacy question focuses on abstract knowledge, such as mathematics; another focuses on language. Essentially, the adequacy questions lead into the basic programmatic adequacy of interactivism in general, and the answers constitute a general model of cognition, perception, representation, and language. These programmatic issues will not be pursued here. What is critical for current purposes is that we have found a conception of representation that is not an encoding, not subject to the many logical incoherencies of encodingism.

The fundamental new property is that interactive representation is functionally emergent in organizations of interactive systems. That is, it constitutes a model of the emergence of representation out of action. As such, it does not fall to the incoherence problem because the representational content is emergent in the strategies and heuristics that are selected, and they do not require any prior representations to come into being or to be used. It does not fall to the origins problem because the differentiation into *A* or *B* does not require the prior knowledge of what *A* or *B* environments are, nor the prior knowledge of which sort of environment the system is currently in. It does not fall to the skeptical problem because the functional information that the system is in, say, an *A* type environment is tautologically certain, while knowledge of *properties* of *A* type environments in the strategies and heuristics are defeasible and can in fact be tested—checked by *using* those strategies and heuristics to actually engage the environment, and checking to see if they work, if they succeed. The fact that representation is *emergent* from action has as one critically important consequence that representation can be checked *via* action without encountering the circularities of skepticism. Without that emergence, checking representation via action gets nowhere because there are no determinate representational interpretations of the actions or of their outcomes: there is no determinate crossing from action back to representation.

Further, interactive representation can serve as the ground for encodings: stand-ins for indicator states like *A* and *B* can be constructed and processed and can be useful for exactly the reasons encodings are useful. A simple differentiator might function strictly passively, although that is intrinsically of reduced power relative to interactive versions, and one form of such a passive differentiator would be a simple sensory transducer, another would be a PDP network. In this perspective, both the potential power of the connectionist approach and the limitations from their intrinsic passivity and non-goal-directedness, are evident: connectionist systems are passive differentiators, and as such, they cannot differentiate what would require *interaction* to differentiate, and they have no representational content—their activation patterns constitute empty "symbols." Interactive representation intrinsically and necessarily involves *open, interactive, goal-directed systems* of exactly the sort discussed by the antirepresentational roboticists: such robots, in the interactive view, *do* in fact involve representation—representation in its most fundamental form as a functional aspect of successful goal-directed interaction—and, therefore, there is a natural bridge to more standard encoding representations *in those circumstances in which encodings would be useful.* The general claim, clearly, is that the interactive model of representation captures the strengths of all three alternative approaches, without encountering their limitations and logical incoherences.

VII. Some Connections and Implications

Interactivism accommodates the correspondences involved in sensory neural activity but with a distinctly nonencoding interpretation: they constitute differentiations that are useful to the further functioning of the overall system in many and various ways, but they do not in any legitimate sense consti-

tute representations of the light patterns, and so on, that they in fact differentiate. The level of analysis concerned with what is in fact differentiated is important to understanding *how* those sensory differentiations manage to be useful to the organism but do not constitute analyses of what the organism knows or represents.

Intermodulations of neural ensemble oscillatory activity *are* control relationships. They are precisely what an interactive control system necessarily involves: a *control* system because that is the level at which notions of interactive system and goal-directed system are constituted and an *oscillatory and modulatory* control system because successful action—thus, successful representation in most cases—requires correct timing at all levels. Interactive representation is emergent not from abstractly sequenced action but from correctly timed *inter*action.

In the interactive view, representation emerges in the organization of the ongoing oscillatory and modulatory activities of the CNS—differentiations of environments and subsequent differentiations, selections, of further activity. There is no need—in fact, it is *in general* quite inappropriate—to attempt to interpret those oscillations and modulations as *themselves* constituting representations (with the caveat of derivative, secondary encodings specialized for and based on that emergent representational function).

In particular, there is no need to attempt to understand language activity in terms of various encodings being transmitted from homunculus to homunculus in the brain for various processings and understandings. In the context of the currently dominant encoding understanding of the nature of language, it has been difficult to avoid this form of analysis of language phenomena at the level of neural functioning. In fact, language *cannot* be fundamentally an encoding phenomena for exactly the same reasons that perception cannot be: encodings cannot provide representational content that is not already there, including knowledge of what an utterance or a word is supposed to represent. Thus, it would be impossible to either learn or to understand utterances if language were in fact merely encodings. Wittgenstein, among others, made essentially this point some time ago, but, in the absence of alternative conceptions and the dominance of the information processing approach in general, it has had limited impact. [*See* LANGUAGE.]

VIII. Summary

Studies of cognitive phenomena in the brain have tended to maintain the same presuppositions concerning the fundamental nature of cognition and representation as has cognitive psychology in general. For some decades, those presuppositions were massively dominated by the information processing view, so much so that it began to take on a sense of taken-for-granted obviousness. More recently, several alternatives to the information processing view, and critiques of that view, have emerged. An unexamined taken-for-grantedness concerning cognition and prepresentation thus no longer suffices.

I have reviewed four of these views and argued that three of them—information processing, connectionism, and a version of robotics—although all interesting and different from each other in important ways, nevertheless share an underlying assumption concerning the nature of representation—an encodingist assumption. Furthermore, this encodingist assumption is wrong and logically incoherent in its foundations.

An alternative interactive model of representation is outlined, and it is argued that it captures the strengths of each of the other three appraoches and avoids their limitations and logical weaknesses. This approach introduces new understandings of the nature and significance of sensory and CNS activity and gives rise to novel questions concerning, and approaches to, such phenomena as language.

Bibliography

Baars, B. J. (1986). "The Cognitive Revolution in Psychology." Guilford, New York.

Bickhard, M. H. (1980). "Cognition, Convention, and Communication." Praeger, New York.

Bickhard, M. H. (1987). The social nature of the functional nature of language, *in* "Social and Functional Approaches to Language and Thought (M. Hickmann, ed), pp. 39–65. Academic Press, New York.

Bickhard, M. H. (in press). The import of Fodor's anticonstructivist arguments, *in* "Epistemological Foundations of Mathematical Experience" (L. Steffe, ed.). Springer-Verlag, New York.

Bickhard, M. H., and Richie, D. M. (1983). "On the Nature of Representation: A Case Study of James J. Gibson's Theory of Perception." Praeger, New York.

Brooks, R. A. (1987). "Intelligence without Representation." MIT Artificial Intelligence Laboratory, manuscript.

Burnyeat, M. (1983). "The Skeptical Tradition." University of California Press, Berkeley.
Campbell, R. L., and Bickhard, M. H. (1986). "Knowing Levels and Developmental Stages." Karger, Basel.
Carlson, N. R. (1986). "Physiology of Behavior." Allyn and Bacon, Boston.
Chapman, D., and Agre, P. (1986). Abstract reasoning as emergent from concrete activity, *in* "Reasoning about Actions and Plans, Proceedings of the 1986 Workshop" (M. P. Georgeff and A. L. Lansky, eds.). Timberline, Oregon, 411–424.
Gardner, H. (1987). "The Mind's New Science." Basic Books, New York.
Glass, A. L., Holyoak, K. J., and Santa, J. L. (1979). "Cognition." Addison-Wesley, Reading Mass.
Kenny, A. (1973). "Wittgenstein." Harvard Univ. Press, Cambridge.
McClelland, J., and Rumelhart, D. (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2. Psychological and Biological Models." MIT Press, Cambridge.
Neisser, U. (1967). "Cognitive Psychology." Appleton-Century-Crofts, New York.
Palmer, S. E. (1978). Fundamental aspects of cognitive representation, *in* "Cognition and Categorization" (E. Rosch and B. B. Lloyd, eds.). Erlbaum, Hillsdale, N.J.
Pinker, S., and Mehler, J. (1988). "Connections and Symbols." MIT Press, Cambridge.
Rumelhart, D., and McClelland, J. (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. Foundations." MIT Press, Cambridge.
Thatcher, R. W., and John, E. R. (1977). "Functional Neuroscience, Vol. 1. Foundations of Cognitive Processes." Erlbaum, Hillsdale, N.J.