CogSci News

Cognitive Science Program, Lehigh University, Bethlehem, PA 18015

Volume 4, Number 1
Spring 1991

Editorial Policy

This newsletter is published twice each year, in fall and spring issues, by the Cognitive Science Program at Lehigh University. Its purpose is to inform faculty and students about the interdisciplinary and rapidly growing field of cognitive science and to report the activities of Lehigh's Program.

The newsletter is distributed free of charge in the United States and Canada to academic programs and individuals interested in cognitive science. Anyone who would like to be added to the mailing list should notify the Editor.

The Editorial Staff welcomes readers' comments and *solicits materials* dealing with cognitive science. We are especially pleased to consider course syllabi, book reviews, short essays, brief descriptions of scholarship and research in progress, and original art work (e.g., cartoons, line-drawings, computer-generated images).

Address all submissions, comments, and subscription requests to: John B. Gatewood, CogSci News, Price Hall #40, Lehigh University, Bethlehem, PA 18015. Send electronic mail to jbgl@lehigh.bitnet or jbgl@ns.cc.lehigh.edu.

How to Build a Machine with Emergent Representational Content

Mark H. Bickhard
Henry R. Luce Professor of Cognitive Robotics
and the Philosophy of Knowledge
Lehigh University

ABSTRACT:

Encodingism as a *fundamental* model of the nature of representation rests on a logical incoherence. This incoherence is manifest, among many other ways, in the empty symbol problem—the inability to provide any representational content to the symbols upon which cognitive science depends. Interactivism is a model of representation that avoids the incoherences and aporias of encodingism—in fact, it explains them. It provides a model of the emergence of functional representation out of non-representational phenomena—in fact, out of relatively simple principles of interactive system organization. It also provides a model for the emergence of encodings out of an interactivist representational foundation. It thereby provides a perspective within which human representation can be modeled and understood. *And* it thereby provides an approach within which machines with emergent representational content can be built.

It is by now a commonplace that cognitive science does not know how to provide its representations with representational content. One generic term for this embarrassment is "the empty symbol problem" (Block 1980, 1981; Harnad 1987; Haugeland 1981)—the sense in which the so- called symbols of cognitive science are not really symbolic of anything—they are empty of representational content—except, of course, for the designer or user or observer. In other words, the problem is to provide representational content *for the system or machine itself.*

The first task is a diagnosis: Why is it so difficult to provide representational content? The answer, simply, is that the field labors under a false and ultimately incoherent conception of the nature of representation, one that makes any such non-empty, genuine representation *impossible.* The second task is to offer a solution: a model of the nature of representation that makes it possible to design genuine representations into a machine—or to understand the basic nature of genuine representations in human beings. Along the way, I will show the sense in which, for all of its differences, connectionism does not offer a solution to this basic problem.

Encodingism: The Problem

The problem, simply, is the almost universal assumption that the nature of representation is that of *encodings* (e.g., Newell 1980). Not in the sense that encodings don't actually exist—they clearly do, and are quite, even essentially, useful. The problem, I argue, is that encodings are inherently a subsidiary, a derivative, form of representation. They intrinsically depend on a more foundational form of representation, and cannot exist without it. To assume that encodings *are* the nature of representation, then—to assume that they are primary and foundational rather than derivative—is to require that they *serve*

(continued on page 2)

Emergent Content (cont.)

the foundational representational function that they in fact only *presuppose* and are derived from. That is, to assume that encodings are the nature of representation is to engage in a deep and inherent circularity. This circularity, in turn, is what makes standard approaches to designing or modeling representational content impossible.

The first step in understanding this circularity is an explication of the nature of encodings. I discuss three equivalent characterizations of the nature of encodings. Note that the general notion of encoding subsumes many distinctions common within the artificial intelligence and cognitive science literature, such as between scripts, frames, and semantic nets, and even the higher level distinction between symbolic and non-symbolic encodings.

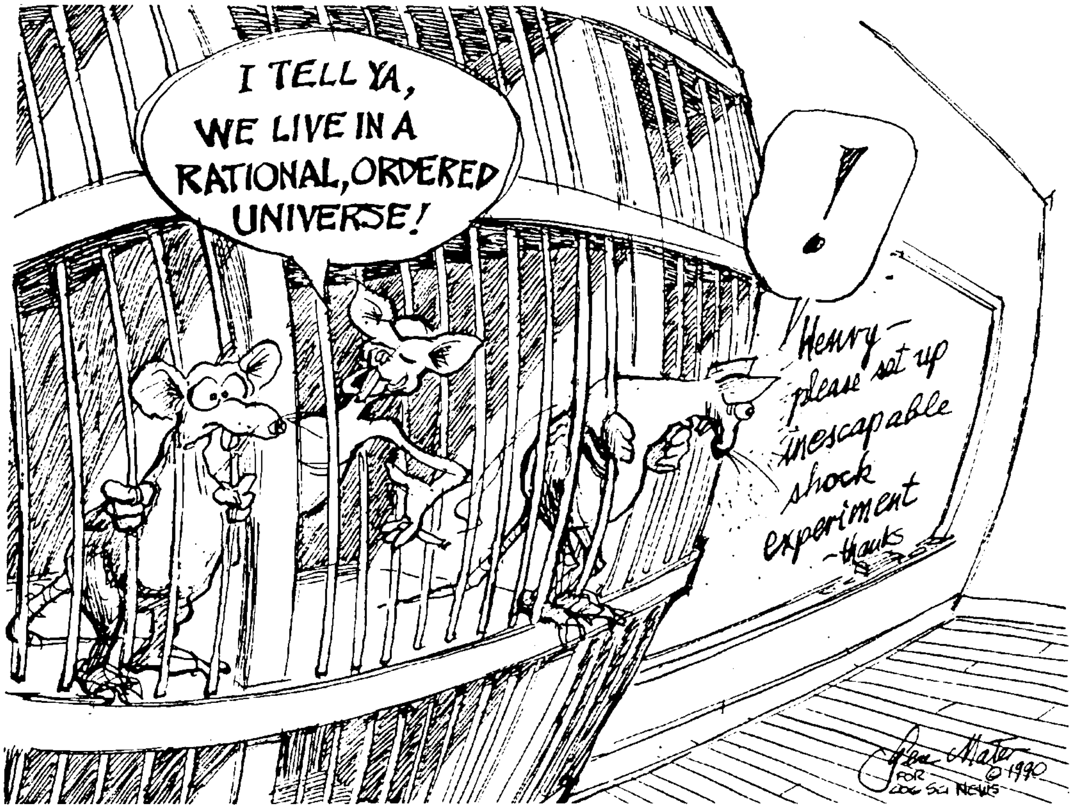
Stand-ins. The first characterization is the most paradigmatic in understanding the actual nature of encodings, though it is not the most seductive with respect to attempting to understand the nature of representation more broadly. This is the sense in which encodings are representational stand-ins. Encodings stand-in for some other representation. Morse code, or any equivalent, provides basic examples: "..." stands-in for "S", while "---" stands-in for "O".

There is nothing exceptionable about this stand-in form of encoding. It is, in fact, the underlying nature of all genuine encodings, although this is easily obscured and not understood. It is precisely this stand-in character of encodings that makes them so useful: encodings change the form and the medium of representation, thereby allowing things to be done to them and with them that would be impossible otherwise. "..." can be sent over telegraph lines, while "S" cannot, and the speeds of manipulation and density of storage attainable with bit patterns are unimaginable with a paper medium.

The stand-in perspective on encodings makes their intrinsic and derivative nature explicitly clear. As stand-ins, they require the prior existence of that which is to be stood-in-for. They require something else to provide the representational content that they carry. These stood-in-for providers of representational content, of course, can themselves be encodings, but those encodings will in turn be stand-ins for some other representations. This iteration of stand-in relationships can continue for some time, but only finitely many times: there must be some ground of representation that provides representational content to the whole hierarchy of encoding stand-ins without itself borrowing that content from—standing-in for—still another layer of representations. That is, there must be some layer of logically independent, non-derivative, non-stand-in, representations. It is only when this grounding level of representation is itself assumed or presupposed to be constituted as encodings that we encounter the circular incoherence of encodingism as an approach to representation.

The circularity is simple: encodings must *borrow* their representational content, while this grounding level must account for the *emergence* of representation. It must account for the emergence of representation out of non-representational

(continued on page 3)


I TELL YA, WE LIVE IN A RATIONAL, ORDERED UNIVERSE!
!
Henry— please set up inescapable shock experiment —thanks
Gene Mater ©1990 FOR COG SCI NEWS

by Gene Mater (Bethlehem, PA)

Emergent Content (cont.)

phenomena because if it derives its representational content from any other representational phenomena, then it is not logically independent, but is simply another layer of derivative stand-ins.

If we assume that this grounding level is constituted as encodings, then we get, for some purported grounding "**X**", " '**X**' represents whatever it is that '**X**' represents." The circularity here is blatant and unavoidable— any other way of attempting to provide representational content would depend on some other already available representation, contrary to the assumption of logical independence. The circularity, in fact, provides no representational content whatsoever, and, thus, fails to make "**X**" an encoding at all. The impossibility of doing this makes the notion of a logically independent, grounding, encoding an *incoherent concept.* The concept presupposes that encodings can generate and provide emergent representational content, but they cannot because they presuppose it.

Shifting perspective slightly, I note that representation clearly *has* emerged out of non-representational content at some point or points in the history of the universe—if not continuously in the learning and development of individuals of many species. Encodings cannot account for such emergence, and, therefore, intrinsically cannot constitute a valid characterization of the fundamental nature of representation. Encodingism implies that representation cannot come into existence, and, therefore, cannot exist at all—a simple reductio ad absurdum. There must be some other form of representation not subject to this incoherence.

Other perspectives on encodings tend to obscure their intrinsic stand-in nature, and, thus, to obscure the above intrinsic incoherence. Harnad (1987), for example, presents a critique similar to the incoherence argument, but seems to assume that its force is limited to symbolic representations. But these other perspectives are all either metaphoric extensions of the notion of encoding that are not epistemic or representational at all—for example, the control-system functional selectivity that leads us to write of genes "encoding" proteins—or they are in fact logically equivalent to the stand-in characterization. I look at two additional such perspectives.

Known Content. The first is the phenomenological paradigm of encoding: it is the notion of a representation as being constituted by something being in a known representational relationship with what it represents. We know, for example, what some map symbol represents, and it is precisely our knowledge that makes it a representation for us. As above, this is unexceptionable in itself, *so long as the implicit stand-in relationship involved is not obscured.* In this case, we must know *that* the map element does represent, and we must know *what* it represents, in order for it itself to *be* a representation for us. That is, we must know that there is a stand-in relationship, and we must know what the representational content is that is being borrowed. Whatever representation it is that is specifying what the map symbol represents is what the map symbol is standing-in-for. The sense in which the stand-in relationship is more obscure in such cases is that the representation being stood-in-for may be more subjective— in the mind of the map reader—rather than explicitly externalized as with the dots, dashes, and characters of Morse code. Even though more obscure in this sense, however, the stand-in nature of such encodings is still fundamental.

Correspondences. The second superficially non-stand-in perspective on encodings that I will address is that of encodings as correspondences. These correspondences may be factual and lawful in nature, or, more generally, informational, or they may be arbitrary, as in the case of Morse code correspondences. The factual and lawful case is very seductive for encodingism: it seems to remove the arbitrariness of encodings and model them as inherent in the natural lawful processes. The most common version of this is sensory transduction.

Correspondences, however, both lawful and informational and otherwise, are plentiful. They are as ubiquitous as any processes exhibiting the regularities of the laws of physics or chemistry, or any other level of lawful process. Clearly, correspondence *per se* does not make for an encoding. In fact, what makes an encoding is a *known* such correspondence. The fact of such a correspondence, when discovered and known in its own right, and the knowledge of what the correspondence is with, may then provide grounds for taking the element that is in such a correspondence as an encoding of what it is in correspondence with. But what it is in correspondence with *must be already known* before it can be taken as an encoding of that. Whatever representation or representations are involved in knowing what the correspondence is with, then, are the representations that are stood-in for by the encoding element or event. This version too is just another perspective on encodings as representational stand-ins.

This is not to deny that factual correspondences between internal events and external phenomena might be important to system functioning, and, perhaps, even to phenomena of representation. It is to deny, however, that that importance can be captured with the notion of encoding, even with supposedly "transduced" encoding. Transduction, after all, is technically a term for changes in the form of *energy.* The usage of the term extended to the presumed creation of encodings skips over the problems of how and in what sense such energy transductions can generate *representations*—of how and in what sense the system can take the internal events as representations at all, and how the system can know what they are to be taken as representations of. In other words, the notion of transduction fails to address the nature and origin of the presumed representational content of the presumed transduced encodings, and any attempt to do so directly encounters the incoherence problem (Bickhard and Richie 1983). Making sense of the functionality of such factual correspondences as sensory transductions, then, remains to be done. It just cannot be done with the notion of encodings.

Connectionism. A new version of the representation as correspondence view is provided by connectionism. In connectionism, the correspondences are between patterns of activation of nodes in a network and various environmental categories. Furthermore, connectionist, or parallel distributed processing, correspondences are neither lawful in the sense of transductions, nor arbitrary in the sense of symbols, but instead are emergent and trained. They have a number of strengths relative to symbolic encoding processing systems, and also relative weaknesses

(continued on page 4)

(Bickhard and Terveen in prep.; Graubard 1988; Horgan and Tienson 1988; Pinker and Mehler 1988), but, nevertheless, they do not solve the problem of representational content. They provide at best correspondences, not known correspondences—except to the user or designer or observer. There is no representation for the system itself.

Encodings as stand-ins, as representations constituted in terms of having known representational content, and as presumably constituted by correspondences are all three—when they are genuinely encodings at all—variations of each other. In particular, they are all variations of the stand-in encoding, and, therefore, are all subject to the same foundational incoherence and impossibility of emergence if taken as constitutive of the nature of representation.

The impossibility of emergence, and the resultant incoherence, of encodingism are the core failures of the approach. There are, however, *many* derivative consequences, variations, partial insights, and reactions in the literature. I will mention a few.

Innatism. First is the familiar argument for a radical innatism of representation (Bickhard in press-b; Fodor 1975, 1981b, 1983). The argument is a partial recognition of the impossibility of emergent encodings. The conclusion is that learning cannot create new representation, therefore it must all be innate. The failure of emergence of encodings, however, is logical, not a failure of the processes of learning or development, and evolution can no more solve it than can psychology.

Methodological Solipsism. A different run around the circular incoherence of encodingism yields an argument for methodological solipsism (Fodor 1981a). Here, encodings are defined in terms of what they represent. But that implies that our knowledge of what is represented is dependent on knowledge of the world, which, in turn, is dependent on our knowledge of physics and chemistry. Therefore, we cannot have an epistemology until physics and chemistry are finished so that we know what is being represented.

This, however, contains a basic internal contradiction: we have to know what is being represented in order to have representations, but we can't know what is being represented until physics and chemistry are historically finished with their investigations. Fodor concludes that we have a methodological solipsism—that we can only model systems with empty formal symbols until that millennium arrives. But how do *actual* representations work? We can't have actual representations until we know what is to be represented—but to know what is to be represented awaits millennial physics—but physics cannot even *begin* until we have some sort of representations of the world. We have to already have representation before we can get representation. Fodor's conclusion is just a historically strung out version of the incoherence problem—another reductio ad absurdum disguised as a valid conclusion about psychology and epistemology.

Constructive Circularity. The incoherence problem focuses on the impossibility of specifying what a logically independent encoding is supposed to represent—of providing any representational content. If we ask instead how we are—or any other system is—to know what encodings to set up, to create, in the first place, either developmentally or perceptually, we encounter a different version of the incoherence. In particular, in order to know what representations to create, we must first know what it is that we are to represent—we must first know what the world is like. But such representations are supposedly the only epistemic means available to us for knowing what the world is like. We have to already have our encodings of the world in order to create our encodings of the world (Piaget 1970).

Skepticism. Still another possible question is that of how we can check our representations for accuracy. The only epistemic access we have to the world is in terms of our encodings; therefore, the only way we can check our encodings is against our encodings. This is a different manifestation of the incoherent circularity of encodingism, and gives no ground whatsoever for any purported accuracy of our representations of the world.

This version of the incoherence is, in its many forms, the historical problem of skepticism (Annas and Barnes 1985; Burnyeat 1983; Stroud 1984; Popkin 1979; Rescher 1980). It has never been solved, but, since its consequences are so evidently absurd—that we do not have any valid representations of the world—it is generally ignored. I suggest that the basic argument is perfectly valid, except that it only applies to encoding representations. Therefore, the proper conclusion is that we do not have any primary encoding representations of the world.

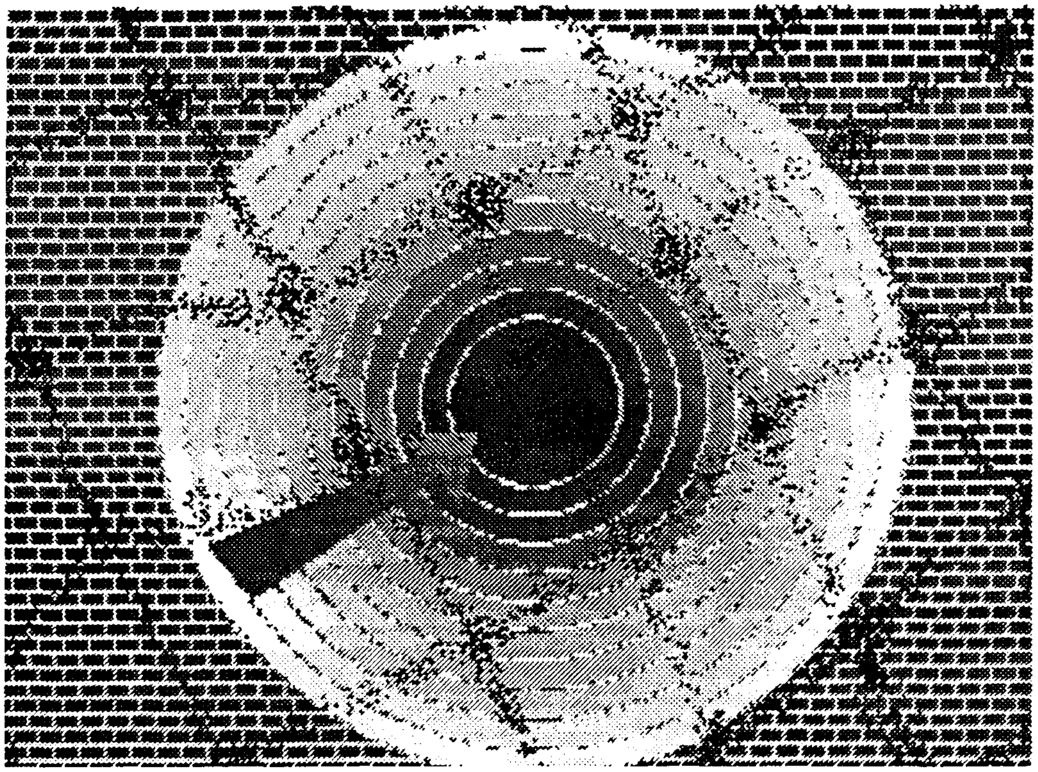
A frequent despairing response to the problem of skepticism has been to conclude that we cannot in fact epistemically get outside of our basic representations of the world. Therefore, our world *is* just those representations and no more. To postulate beyond that is superfluous and invalid. A version of this at the level of the individual is classical solipsism or idealism—the world is just my dream. A more sophisticated and more common version in contemporary literature is to make the argument at the level of language representations, yielding a linguistic idealism (Bickhard 1987; Bickhard and Terveen in prep.; Gadamer 1975; Heidegger 1962; Maturana and Varela 1980, 1987). Idealisms, however, are just encodingisms without any encoded world—'bare' representational content that represents only itself. Arguments for idealisms fail along with the failure of encodingism in general. In any case, idealisms provide no more solution to the basic incoherence of how we are supposed to know what is supposed to be being represented in the first place than does any other version of encodingism. Idealisms give up on the question of accuracy with respect to an external world by jettisoning the external world, but do not address the incoherence nor the emergence problems.

Basically, encodingism presents an infinitude of blind alleys. Many of these are exemplified in historical and contemporary literature; many undoubtably remain to be discovered. If the basic critique of encodingism is correct, however, then there is *nothing but* such blind alleys within the encodingism approach. Furthermore, it should be very carefully noted that it is impossible to discover that you are trapped in an infinity of blind alleys by exploring them one—or a whole bunch—at a time. This is true at the historical level, at the individual level, and at the level of cognitive science as a collective endeavor.

Interactivism: The Solution

Interactivism provides a *functional* model of representation. That is, it presents a functional explication of representat*ion*, rather than a characterization of representat*ions*. Any representation, in fact, *is* a representation for any epistemic system only insofar as it *functions appropriately* for that system—whatever such appropriate functioning might be. Conversely, anything that does function appropriately for a system will by virtue of

(continued on page 5)



(computerized image by Annie Paghidas)

Emergent Content (cont.)

that be a representation, or serve the function of representation, for that system.

It is possible that, in order for an element or a structure to be *able* to participate in a representational function for a system, there would be further constraints on the nature or origin or organization of that element or structure. There is, in fact, a degree of trade-off here, with some systems relying more on particular properties of their representations, and others presupposing extremely little about the instantiations of representations. Even here, however, the issue is fundamentally at the level of the functioning of the system, with any presuppositions about, or constraints on, the representations being derivative from such functioning. Contrary to simple encodingism, representation is fundamentally functional in nature, not structural.

This relatively simple point already yields a new perspective on the incoherence problem: an encoding serves as a representation for a system insofar as the system makes use of it as a representation—makes use of it as carrying representational content. But, the ability of the system to make use of it as carrying representational content *constitutes its having* that representational content. In other words, an encoding's having representational content is a property of the functional usage of the encoding by the system—it is a property of the system knowing what the encoding is supposed to represent—and not a property of the encoding element itself. To presuppose, then, that an encoding can provide its own representational content—can be other than a representational stand-in—is to presuppose that it can somehow carry or accomplish its own representational functional usage. But an encoding *element* qua encoding *element* is not a system at all, and "functional" is a system-relational concept—an element *cannot* have a function except relative to something other than itself, relative to some system.

In the broadest sense, the only function that a representation could serve internal to a system is to select, to differentiate, the system's further internal activities. This is the basic locus of representational function, but there are two additional logical necessities. The first is that the differentiation of system activities be in some sense epistemically related to some environment being represented. The second is that those differentiations in some sense constitute at least implicit predications that could be wrong *from the perspective of the system itself.* (Simply being wrong per se allows any observer semantics to determine such 'wrongness' and thus yields a semantics for that observer, but not for the system itself.)

Differentiations. Consider an interactive system engaged in interaction with its environment. The internal course of the interactive process will in part depend on the system organization, and in part on the environment being interacted with. The internal state that the system is in when the interaction is finished will, thus, serve to differentiate those environments that would yield that final internal state from those environments that would yield some other final state. The possible internal final states of an interactive system or subsystem, then, serve as potential differentiators of the environments that would yield them. Those potential final states *implicitly define*—in an interactive generalization of model theoretic implicit definition (Bickhard 1980a; Campbell and Bickhard 1986)—the differentiated categories of environments.

Note first that such interactive differentiation is *of the environment.* It constitutes a form of epistemic contact with the environment. Note second that this epistemic contact is not itself an encoding contact—there is no representational content *about* those differentiated, those implicitly defined, classes of environments. There is no

(continued on page 6)

representational content that could make those final states into encodings. Note third that such a relationship of differentiation, of implicit definition, will constitute *factual correspondences* with whatever is thereby differentiated. Passive versions of such differentiations, therefore factual correspondences, are exactly what are established by, for example, *transduction.* A different form of such passive differentiations, therefore factual correspondences, are exactly what are established by connectionist or PDP systems. In the interactive case, however, there is no claim or presumption that the system has any representational content about what is being differentiated merely by virtue of having made the differentiation. It is exactly this latter invalid conclusion that yields the epistemic encoding interpretation of transduction and of connectionist systems.

At this point in the analysis, we have differentiations, that have no representational content, and, therefore are not and cannot be encodings. But we do not have representational content—no implicit predications about those differentiated environments that could be right or wrong about those environments.

Content. Consider next an interactive differentiator embedded as a subsystem in a larger goal-directed interactive system. Suppose that the possible final states of the differentiator are, say, **A** and **B**. Suppose further that the system organization is such that, when, say, internal goal **G202** is active, and differentiating final state **A** has been reached, then the system executes strategy **S27**, while if final state **B** has been reached, then the system executes strategy **S433**. In this configuration, the final states serve the function of selecting, of differentiating, of *indicating* the relevant potentiality for, the further activity of the system in the anticipated service of the goal. This is the locus of the representational function.

Such indications constitute implicit predications about final state **A** type environments and final state **B** type environments. In particular, they predicate that "**A** type environments are strategy **S27** type environments" and "**B** type environments are **S433** type environments". That is, that **A** type environments have the interactive properties appropriate to strategy **S27**, and so on.

Furthermore, these predications could be wrong. Indicated strategies might not work—might not yield interactions that can be controlled by those strategies—at all. Still further, that wrongness is definable, and potentially detectable, from within the system itself. The strategies attempt control of the interactions in the service of the relevant goals, and goal failures constitute *falsification* of the general indicative predication. This is the fundamental point of emergence of representational content. It is an emergence of representational function out of system functional organization that is not itself already necessarily representational.

It is also a function for which specialized elements and machinery could be designed and constructed—derivative encodings (Bickhard and Richie 1983). Such encodings would always, intrinsically, be dependent upon the interactive representational emergence for the representational content that makes them encodings.

Neglected Issues. Many issues concerning interactive representation have been neglected here. The epistemic *asymmetry* between the correctness and incorrectness of the interactive representational indications (Bickhard 1980a, in press-a); the inherent *modality* of such interactive representation—unlike the pure extensional actuality focus of most encoding systems (Bickhard 1980a, 1988a, 1988b; Bickhard and Campbell 1989); the sense in which functional relevance *is* functional representational indication (Bickhard 1980a; Bickhard and Terveen in prep.); the significance of the *splitting* of epistemic contact in differentiation from representational content in further interactive indications—unlike encodings for which epistemic contact *is* the carrying of representational content (Bickhard and Richie 1983); the sense in which the emergence out of action and interaction provides a *non-circular check* on representation (Bickhard 1987, in press-c); and so on. The basic concern has been simply to show that an alternative to encodingism is required, and that interactivism satisfies at least minimal requirements for such an alternative. In particular, that it solves the problem of emergence—and, therefore, avoids the incoherence problem and related difficulties.

I have also not addressed at all the sometimes radical revisions that interactivism requires in approaches to general cognitive phenomena. Perception cannot be an encoding of the environment (Bickhard and Richie 1983); language cannot be an encoding of mental contents (Bickhard 1980a, 1987; Bickhard and Terveen in prep.); knowing, learning, emotions, and consciousness acquire system specific models (Bickhard 1980b; Campbell and Bickhard 1986); neither Turing machine theory nor Tarskian model theory *nor any of their equivalents* are adequate mathematical grounds for cognitive science (Bickhard and Terveen in prep.); arguments for innatism, modularity, inherent cognitive limitation, methodological solipsism, and others are undercut (Bickhard in press-b; Bickhard and Richie 1983; Bickhard and Terveen in prep.); genuine development *can* occur (Bickhard in press-b; Campbell and Bickhard 1986); novel phenomena, constraints, and forms of explanation emerge (Bickhard in press-c); and on and on (Bickhard and Campbell in prep.). Basically, representation is everywhere, and interactivism radically revises our notions of representation. On the other hand, interactive representation *is* emergent, and is so out of principles of system organization that are *in-principle* easy to design and to build. In practice, however, interesting interactive representation will be quite complex (Bickhard 1980a).

Interactivism avoids the incoherences and aporias of encodingism—in fact, it explains them. It provides a model of the emergence of functional representation out of non-representational phenomena—in fact, out of relatively simple principles of interactive system organization. It thereby provides a perspective within which human representation can be modeled and understood. *And* it thereby provides an approach within which machines with emergent representational content can be built.

References

Annas, J. and Barnes, J. (1985) *The Modes of Scepticism.* New York: Cambridge University Press.

Bickhard, M. H. (1980a) *Cognition, Convention, and Communication.* New York: Praeger.

Bickhard, M. H. (1980b) A model of developmental and psychological processes. *Genetic Psychology Monographs* 102:61-116.

Bickhard, M. H. (1987) The social nature of the functional nature of language. In M. Hickmann, ed., *Social and Functional Approaches to Language and Thought.* New York: Academic Press.

(continued on page 8)