FOUNDATIONAL ISSUES IN ARTIFICIAL INTELLIGENCE AND COGNITIVE SCIENCE

IMPASSE AND SOLUTION

Mark H. Bickhard
Lehigh University

Loren Terveen
AT&T Bell Laboratories

forthcoming, 1995

Elsevier Science Publishers

Contents

Preface

Artificial Intelligence and Cognitive Science are at a foundational impasse which is at best only partially recognized. This impasse has to do with assumptions concerning the nature of representation: standard approaches to representation are at root circular and incoherent. In particular, Artificial Intelligence research and Cognitive Science are conceptualized within a framework that assumes that cognitive processes can be modeled in terms of manipulations of encoded symbols. Furthermore, the more recent developments of connectionism and Parallel Distributed Processing, even though the issue of manipulation is contentious, share the basic assumption concerning the encoding nature of representation. In all varieties of these approaches, representation is construed as some form of encoding correspondence. The presupposition that representation is constituted as encodings, while innocuous for *some **applied*** Artificial Intelligence research, is fatal for the further reaching programmatic aspirations of both Artificial Intelligence and Cognitive Science.

First, this encodingist assumption constitutes a *presupposition* about a basic aspect of mental phenomena — representation — rather than constituting a *model* of that phenomenon. Aspirations of Artificial Intelligence and Cognitive Science to provide any foundational account of representation are thus doomed to circularity: the encodingist approach presupposes what it purports to be (programmatically) able to explain. Second, the encoding assumption is not only itself in need of explication and modeling, but, even more critically, the standard presupposition that representation is *essentially* constituted as encodings is logically fatally flawed. This flaw yields numerous subsidiary consequences, both conceptual and applied.

This book began as an article attempting to lay out this basic critique at the programmatic level. Terveen suggested that it would be more powerful to supplement the general critique with explorations of actual projects and positions in the fields, showing how the foundational flaws visit themselves upon the efforts of researchers. We began that task, and, among other things, discovered that there is no natural closure

to it — there are always more positions that could be considered, and they increase in number exponentially with time. There is no intent and no need, however, for our survey to be exhaustive. It is primarily illustrative and demonstrative of the problems that emerge from the underlying programmatic flaw. Our selections of what to include in the survey have had roughly three criteria. We favored: 1) major and well known work, 2) positions that illustrate interesting deleterious consequences of the encodingism framework, and 3) positions that illustrate the existence and power of moves in the direction of the alternative framework that we propose. We have ended up, *en passant*, with a representative survey of much of the field. Nevertheless, there remain many more positions and research projects that we would like to have been able to address.

The book has gestated and grown over several years. Thanks are due to many people who have contributed to its development, with multitudinous comments, criticisms, discussions, and suggestions on both the manuscript and the ideas behind it. These include, Gordon Bearn, Lesley Bickhard, Don Campbell, Robert Campbell, Bill Clancey, Bob Cooper, Eric Dietrich, Carol Feldman, Ken Ford, Charles Guignon, Cliff Hooker, Norm Melchert, Benny Shanon, Peter Slezak, and Tim Smithers. Deepest thanks are also due to the Henry R. Luce Foundation for support to Mark Bickhard during the final years of this project.

Mark H. Bickhard
Henry R. Luce Professor of
Cognitive Robotics & the Philosophy of Knowledge
Department of Psychology
17 Memorial Drive East
Lehigh University
Bethlehem, PA 18015
mhb0@lehigh.edu

Loren Terveen
Human Computer Interface Research
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
terveen@research.att.com

Introduction

How can we understand representation? How can we understand the mental? How can we build systems with genuine representation, with genuine mentality? These questions frame the ultimate programmatic aims of Artificial Intelligence and Cognitive Science. We argue that Artificial Intelligence and Cognitive Science are in the midst of a programmatic impasse — an impasse that makes these aims impossible — and we outline an alternative approach that transcends that impasse.

Most contemporary research in Artificial Intelligence and Cognitive Science proceeds within a common conceptual framework that is grounded on two fundamental assumptions: 1) the unproblematic nature of formal systems, and 2) the unproblematic nature of encoded, semantic symbols upon which those systems operate. The paradigmatic conceptual case, as well as the paradigmatic outcome of research, is a computer program that manipulates and operates on structures of encoded data — or, at least, a potentially programmable model of some phenomena of interest. The formal mathematical underpinnings of this approach stem from the introduction of Tarskian model theory and Turing machine theory in the 1930s. Current research focuses on the advances to be made, both conceptually and practically, through improvements in the programs and models and in the organization of the data structures.

In spite of the importance and power of this approach, we wish to argue that it is an intrinsically limited approach, and that these limits not only fall far short of the ultimate programmatic aspirations of the field, but severely limit some of the current practical aspirations as well. In this book, we will explore these limitations through diverse domains and applications. We will emphasize unrecognized and unacknowledged programmatic distortions and failures, as well as partial recognitions of, and partial solutions to, the basic impasse of the field. We also slip in a few additional editorial comments where it seems appropriate. In the course of these analyses, we survey a major portion of contemporary Artificial Intelligence and Cognitive Science.

The primary contemporary alternative to the dominant symbol manipulation approach is connectionism. It might be thought to escape our critique. Although this approach presents both intriguing differences and strengths, we show that, in the end, it shares in precisely the fundamental error of the symbol manipulation approach. It forms, therefore, a different facet of the same impasse.

The focus of our critique — the source of the basic programmatic impasse — is the assumption that representation is constituted as some form of *encoding*. We shall explicate what we mean by "encoding" representation and show that Artificial Intelligence and Cognitive Science universally presupposes that representation *is* encoding. We argue that this assumption is logically incoherent, and that, although this incoherence is innocuous for some purposes, — including some very useful purposes — it is fatal for the programmatic aspirations of the field.

There are a large number of variants on this assumption, many not immediately recognizable as such, so we devote considerable effort to tracing some of these variants and demonstrating their equivalence to the core encoding assumption. We also analyze some of the myriads of deleterious consequences in dozens of contemporary approaches and projects. If we are right, the impasse that exists is at best only dimly discerned by the field. Historically, however, this tends to be the case with errors that are programmatic-level rather than simply project-level failures. Many, if not most, of the problems and difficulties that we will analyze are understood as problems by those involved or familiar with them, but they are not in general understood as having any kind of common root — they are not understood as reflecting a general impasse.

We also introduce an alternative conception of representation — we call it *interactivism* — that avoids the fatal problematics of encodingism. We develop interactivism as a contrast to standard approaches, and we explore some of its consequences. In doing so, we touch on current issues, such as the frame problem and language, and we introduce some of interactivism's implications for more powerful architectures. Interactivism serves both as an illuminating contrast to standard conceptions and approaches, and as a way out of the impasse.

A PREVIEW

For the purpose of initial orientation, we adumbrate a few bits of our critique and our alternative. The key defining characteristic of encodingism is the assumption that representations are constituted as

correspondences. That is, there are correspondences between "things-in-the-head" (e.g., states or patterns of activity) of an epistemic agent (e.g., a human being or an intelligent machine) — the encodings — and things in the world. And, crucially, it is *through this correspondence* that things in the world are represented. It is generally understood that this is not sufficient — there are too many factual correspondences in the universe, and certainly most of them are *not* representations — so much effort is expended in the literature on what additional restrictions must be imposed in order for correspondences to be representational. That is, much effort is devoted to trying to figure out what *kinds* of correspondences are encodings.

One critical problem with this approach concerns how an agent could ever know what was on the other end of a correspondence — *any* correspondence, of *any* kind. The mere fact that a certain correspondence exists is not sufficient. No element in such a correspondence, of any kind, *announces* that it is in a correspondence and what it corresponds to. And we shall argue that so long as our modeling vocabulary is restricted to such factual correspondences, there is no way to provide (to an agent) knowledge of what the correspondences are with. It is crucial to realize that *knowing that* something is in a correspondence and *knowing what* it corresponds to is precisely one version of the general problem of representation we are trying to solve! Thus, as an attempt at explaining representation, encodingism presupposes what it purports to explain.

The interactive alternative that we offer is more akin to classical notions of "knowing how" than to such correspondence-encoding notions of "knowing that." Interactive representation is concerned with functionally realizable knowledge of the potentialities for action in, and interaction with, the world. Interactive representations do not represent what they are in factual correspondence with in the world, but, rather, they represent *potentialities* of interaction *between* the agent and the world. They indicate that, in certain circumstances, a certain course of action is possible. Such potentialities of interaction, in turn, are realizable as the interactive control organizations in the agent that would engage in those interactions should the agent select them.

Obviously, this issues a flurry of promissory notes. Among them are: How is the encodingism critique filled out against the many proposals in the literature for making good on encoding representation? What about the proposals that don't, at least superficially, look like encodingism at all? How is interactive representation realized, without

committing the same circularities as encodingism? How is “knowing that” constituted within an interactive model? How are clear cases of encodings, such as Morse or computer codes, accounted for? What are the implications of such a perspective for related phenomena, such as perception or language? What difference does it all make? We address and elaborate on these, and many other issues, throughout the book.

I

GENERAL CRITIQUE

1

Programmatic Arguments

The basic arguments presented are in-principle arguments against the fundamental programmatic presuppositions of contemporary Artificial Intelligence and Cognitive Science. Although the histories of both fields have involved important in-principle programme-level arguments (for example, those of Chomsky and Minsky & Papert, discussed below), the standard activities within those fields tend to be much more focused and empirical *within* the basic programme. In other words, project-level orientations, rather than programme-level orientations, have prevailed, and the power and importance of programmatic-level in-principle arguments might not be as familiar for some as project-level claims and demonstrations.

The most fundamental point we wish to emphasize is that, if a research programme is intrinsically flawed — as we claim for Artificial Intelligence and Cognitive Science — no amount of strictly project-level work will ever discover that flaw. Some, or even many, projects may in fact fail because of the foundational flaws of the programme, but a project-level focus will always tend to attribute such failures to particulars and details of the individual projects, and will attempt to overcome their shortcomings in new projects *that share exactly the same foundational programmatic flaws*. Flawed programmes can never be refuted empirically.

We critique several specific projects in the course of our discussion, but those critiques simply illustrate the basic programmatic critique, and have no special logical power. Conversely, the enormous space of particular projects, both large and small, that we do *not* address similarly has no logical bearing on the programme-level point, unless it could be claimed that one or more of them constitute a counterexample to the programmatic critique. We mention this in part because in discussion with colleagues, a frequent response to our basic critique has been to

name a series of projects with the question "What about this?" of each one. The question is not unimportant for the purpose of exploring how the programmatic flaws have, or have not, visited their consequences on various particular projects, but, again, except for the possibility of a counterexample argument, it has no bearing on the programmatic critique. Foundational problems can neither be discovered nor understood *just* by examining sequences of specific projects.

CRITIQUES AND QUALIFICATIONS

A potential risk of programmatic critiques is that they can too easily be taken as invalidating, or as claiming to invalidate, all aspects of the critiqued programme without differentiation. In fact, however, a programmatic critique may depend on one or more separable *aspects or parts* of the programme, and an understanding and correction at that level can allow the further pursuit of an even stronger appropriately revised programme. Such revision instead of simple rejection, however, requires not only a demonstration of some fundamental problem at the programmatic level, but also a diagnosis of the grounds and nature of that problem so that the responsible aspects can be separated and corrected. Thus, Chomsky's (1964) critique of the programme of associationistic approaches to language seems to turn on the most central defining characteristics of associationism: there is no satisfactory revision, and the programme has in fact been mostly abandoned. Minsky and Papert's (1969) programmatic level critique of Perceptrons, on the other hand, was taken by many, if not most, as invalidating an entire programmatic approach, without the diagnostic understanding that their most important arguments depended on the then-current Perceptron limitation to two layers. Recognition of the potential of more-than-two-layer systems, as in Parallel Distributed Processing systems, was delayed by this lack of diagnosis of the programmatic flaw. On the other hand, the flaw in *two*-layer Perceptrons would never have been discovered using the project-by-project approach of the time. On still another hand, we will be arguing that contemporary PDP approaches involve their own programmatic level problems.

DIAGNOSES AND SOLUTIONS

Our intent in this critique is to present not only a demonstration of a foundational programmatic level problem in Artificial Intelligence and

Cognitive Science, but also a diagnosis of the location and nature of that problem. Still further, we will be adumbrating, but *only* adumbrating, a programmatic level solution. The implications of our critique, then, are not at all that Artificial Intelligence and Cognitive Science should be abandoned, but, rather, that they require programmatic level revision — even if somewhat radical revision.

We are *not* advocating, as some seem to, an abandonment of attempts to capture intentionality, representationality, and other mental phenomena within a naturalistic framework. The approach that we are advocating is very much within the framework of naturalism. In fact, it yields explicit architectural design principles for intentional, intelligent systems. They just happen to be architectures different from those found in the contemporary literature.

IN-PRINCIPLE ARGUMENTS

Both encodingism and interactivism are programmatic approaches. In both cases, this is a factual point, not a judgement: it is *relevant* to issues of judgement, however, in that the forms of critique appropriate to a programme are quite different than the forms of critique appropriate to a model or theory. In particular, while specific results can refute a model or theory, only in-principle arguments can refute a programme because any empirical refutation of a specific model within a programme only leads to the attempted development of a new model within the same programme. The problem that this creates is that a programme with foundational flaws can never be discovered to *be* flawed simply by examining particular models (and their failures) within that programme. Again, any series of such model-level empirical failures might simply be the predecessors to the correct model — the empirical failures do not impugn the programme, but only the individual models. If the programme has *no* foundational flaws, then continued efforts from within that framework are precisely what is needed.

But if the programme does indeed have foundational flaws, then efforts to test the programme that are restricted to the model level are doomed never to find those flaws — only in-principle arguments can demonstrate those. We dwell on this point rather explicitly because most researchers are not accustomed to such points. After all, programmes are overthrown far less frequently than particular models or theories, and most researchers may well have fruitful entire careers without ever experiencing a programmatic-level shift. Nevertheless, programmes do

fail, and programmes do have foundational flaws, and, so our argument goes, Artificial Intelligence and Cognitive Science have such flaws in their programmatic assumptions. The critique, then, is not that Artificial Intelligence and Cognitive Science are programmatic — that much is simply a fact, and a necessary fact (foundational assumptions cannot be simply avoided!) — the critique is that Artificial Intelligence and Cognitive Science involve *false* programmatic assumptions, and the point of the meta-discussion about programmes is that it requires conceptual-level critique to uncover such false programmatic assumptions. Interactivism, too, is programmatic, and necessarily so. Its contrast with other approaches, so we claim, lies in not making false encodingist presuppositions regarding representation as do standard Artificial Intelligence and Cognitive Science.

2

The Problem of Representation

ENCODINGISM

The fundamental problem with standard Artificial Intelligence and Cognitive Science can be stated simply: they are based on a presupposition of *encoded symbols*. Symbols are instances of various formal symbol *types*, and symbol types are formal "shapes" whose instances can be physically distinguished from each other within whatever physical medium is taken to constitute the material system. Such differentiation of physical instances of formal types constitutes the bridge from the materiality of the representations to the formality of their syntax (Haugeland, 1985). These symbol types — formal shape types — generally consist of character shapes on paper media, and bit patterns in electronic and magnetic media, but can also consist of, for example, patterns of long and short durations in sounds or marks as in Morse code.

Symbols, in turn, are assumed to represent something, to carry some representational content. They may be taken as representing concepts, things or properties or events in the world, and so on.

More broadly, encodings of all kinds are constituted as *being representations* by virtue of their carrying some representational content — by virtue of their being taken to represent *something* in particular. That content, in turn, is usually taken to be constituted or provided by some sort of a correspondence with the "something" that is being represented.[1] For example, in Morse code, "• • •" is interpreted to be a representation of the character or phonetic class **S** — with which it is in Morse-code correspondence. By exact analogy, "standard" Artificial

[1] If what is being represented does not exist, e.g., a unicorn, then such an assumption of representation-by-correspondence is untenable, at least in its simple version: there is nothing for the correspondence relationship to hold with. Whether this turns out to be a merely technical problem, or points to deeper flaws, is a further issue.

Intelligence and Cognitive Science approaches to *mental* (and machine) representation assume that a particular mental state, or pattern of neural activity, or state of a machine, is a representation of, say, a dog. As we argue later, this analogy cannot hold.

We will be arguing that all current conceptions of representation are encoding conceptions, though usually not known explicitly by that name, and are often not recognized as such at all. In fact, there are many different approaches to and conceptions of representation that turn out to be variants of or to presuppose encodingism as capturing the nature of representation. Some approaches to phenomena that are superficially not representational at all nevertheless presuppose an encodingist nature of representation. Some approaches are logically equivalent to encodingism, some imply it, and some have even more subtle presuppositional or motivational connections. Representation is ubiquitous throughout intentionality, and so also, therefore, are assumptions and implicit presuppositions about representation. Encodingism permeates the field. We will examine many examples throughout the following discussions, though these will not by any means constitute an exhaustive treatment — that is simply not possible. The arguments and analyses, however, should enable the reader to extend the critique to unaddressed projects and approaches.

Circularity

It is on the basic assumption that symbols provide and carry representational contents that programmatic Artificial Intelligence and Cognitive Science founder. It is assumed that a symbol represents a particular thing, and that it — the symbol — somehow informs the system of what that symbol is supposed to represent. This is a fatal assumption, in spite of its seeming obviousness — what else could it be, what else could representation *possibly* be?

The first sense in which this assumption is problematic is simply that both Artificial Intelligence and Cognitive Science take the carrying of representational content as a theoretical primitive. It is simply assumed that symbols can provide and carry representational content, and, thus, are *encoded* representations. Representation is rendered in terms of *elements* with representational contents, but there is no model of *how* these elements can carry representational content. Insofar as programmatic Artificial Intelligence and Cognitive Science have aspirations of explicating and modeling all mental phenomena, or even just all cognitive

phenomena, here is an absolutely central case — representation — in which they simply presuppose what they aspire to explain. They presuppose phenomena of representation — symbols having content — in their supposed accounts of cognition and representation. Both fields are programmatically circular (Bickhard, 1982).

Incoherence — The Fundamental Flaw

The second sense in which the encodingism of Artificial Intelligence and Cognitive Science is fatal is that the implicit promissory note in the presupposition of encodingism is logically impossible to cash. Not only do both fields presuppose the phenomena of representation in their encodingism, they presuppose it in a form — representations are essentially constituted as encodings — that is at root logically incoherent. There are a number of approaches to, and consequences of, this fundamental incoherence. We will present several of each.

Recall the definition of an encoded representation: a representational element, or symbol, corresponds to some thing-to-be-represented, and it is a representation by virtue of carrying a representational content specifying that thing-to-be-represented. An encoding is *essentially* a carrier of representational content and cannot exist without some such content to carry, hence the notion of an encoding that does not purport to represent something is nonsense. This problem is not fundamental so long as there *is* some way of *providing* that content for the encoding element to carry. Still further, encodings can certainly *be* providers of representational content for the formation of additional encodings, as when "**S**" is used to provide the content for "• • •" in Morse code. This is a simple and obvious transitive relationship, in which an encoding in one format, say "• • •" in Morse code, can stand in for the letter "**S**," and, by extension, for whatever it is that provided the representational content for "**S**" in the first place. These carrier and stand-in properties of encodings account for the ubiquity and tremendous usefulness of encodings in contemporary life and technology. Encodings change the form or substrate of representations, and thus allow many new manipulations at ever increasing speeds. But they do not even address the foundational issue of where such representational contents can ultimately come from.

Encodings can *carry* representational contents, and already established encodings can *provide* representational contents for the formation of some other encoding, but there is no way within

encodingism per se for those representational contents to ever arise in the first place. There is no account, and — we argue — no account possible, of the *emergence* of representation.

An encoding $\mathbf{X_2}$ can stand in for some other encoding $\mathbf{X_1}$, and $\mathbf{X_1}$ thus provides the representational content that makes $\mathbf{X_2}$ a representation at all. That provider-encoding could in turn be a stand-in for still some other encoding, and so on, but this iteration of the provision of stood-in-for representational content cannot proceed indefinitely: $\mathbf{X_3}$ can stand-in for $\mathbf{X_2}$, which can stand-in for $\mathbf{X_1}$, and so on, only finitely many times — there must be a bottom level.

Consider this bottom level of encodings. In order to constitute these elements as encodings, there must be some way for the basic representational content of these elements to be provided. If we suppose that this bottom-level foundation of logically independent representations — that is:

- representations that don't just stand-in for other representations, and, therefore,
- representations that don't just carry previously provided contents —

is also constituted as encodings, then we encounter a terminal incoherence.

Consider some element **X** of such a purported logically independent, bottom level, foundation of encodings. On the one hand, **X** cannot be provided with representational content by any other representation, or else, contrary to assumption, it will not be logically independent — it will simply be another layer of stand-in encoding. On the other hand, **X** cannot provide its own content. To assume that it could yields "**X** represents whatever it is that **X** represents" or "**X** stands-in for **X**" as the provider and carrier relationship between **X** and itself. This does not succeed in providing **X** with any representational content at all, thus does not succeed in making **X** an encoding at all, and thus constitutes a logical incoherence in the assumption of a foundational encoding.

This incoherence is the fundamental flaw in encodingism, and the ultimate impasse of contemporary Artificial Intelligence and Cognitive Science. Representational content must ultimately emerge in some form other than encodings, which can then provide representational contents for the constitution of *derivative* encodings.

A First Rejoinder

One apparent rejoinder to the above argument would simply claim that the stand-in relationship could be iterated one more time, yielding a foundation of basic encodings that *stand-in for things in the world.* In fact, it might be asked, "What else would you expect representations to do or be?" There are several confusions that are conflated in this "rejoinder." First is an equivocation on the notion of "standing-in-for." The stand-in relationship of encodings is one in which a derivative encoding stands-in for a primary encoding in the sense that the derivative encoding *represents the same thing* as does the primary encoding. For example, the Morse code "• • •" represents whatever it is that "**S**" represents. Therefore, this purported last iteration of the stand-in relationship is an equivocation on the notion of "stand-in": the "thing" in the world isn't being taken as representing anything — it is, instead, that which is to *be* represented — and, therefore, the thing in the world cannot be representationally *stood-in-for*. A supposed mental encoding of a cup, for example, does not represent the same thing that the cup represents — the cup is not a representation at all, and, therefore, the cup *cannot* be representationally stood-in-for. The cup might be representationally *stood-for*, but it cannot be representationally *stood-in-for*.

Second, this purported grounding stand-in relationship cannot be some sort of *physical substitution* stand-in: a "thing" and its representation are simply not the same ontological sort — you cannot do the same things with a representation of **X** that you can with **X** itself. A system *could* have internal states that *functionally track* properties and entities of its environment, for the sake of other functioning in the system. And such functional tracking relationships could be called (functional) stand-in relationships without doing any damage to the meanings of the words. Nevertheless, such a tracking relationship, however much it might be legitimately *called* a "stand-in relationship," is not in itself a *representational* relationship. It is not a representational *stand-in* relationship — the tracking state per se neither *represents* what it tracks (there is no knowledge, no content, of what it tracks), nor does it *represent the same thing as* what it tracks.

The purported grounding stand-in relationship, then — the supposed bottom level encoding stand-in of the element "standing-in" for the cup — simply ***is*** the representational relationship. The relationship of the supposed mental encoding of the cup to that cup is not that of a representational *stand-in* at all, but, rather, that of the representational

relationship itself. The encoding, bottom level or otherwise, "stands-in" for the thing in the world in the sense that it *represents* that thing in the world, and that representational relationship is exactly what was supposed to be accounted for; it is exactly the relationship that we set out to understand and to model in the first place.

The purported explication of representation in terms of grounding stand-ins turns out to be a simple semantic circularity: "representation" is being defined in terms of a usage of "stand-in" that means "representation." Furthermore, the grounding encoding can represent its proper thing-in-the-world only if the relevant epistemic agents *know what it represents*, and they can know what it represents only if they *already* know that which is to be represented. We are right back at the circularity: An encoding of **X** can only be constructed if **X** is already known — otherwise, what is the encoding to be constructed as an encoding *of*? — and **X** can be already known only if there is a representation of **X** already available. In other words, an encoding of **X** can exist only if it is defined in terms of an already existing representation of **X**. Within an encodingism, you must already have basic representations before you can get basic representations. The supposed last iteration of the stand-in relationship, then, appears to avoid the vicious circularity only because of the overlooked equivocation on "stand-in." The relationship between mental representations and things-in-the-world cannot be the same as that between "• • •" and "**S**."

There are, of course, much more sophisticated (and more obscure) versions of this rejoinder in the literature. We discuss a number of them below. Whatever the sophistication (or obscurity), however, as long as the basic notion of representation is taken to be that of an encoding, the fundamental incoherence of encodingism as an approach to representation remains. Strict encodingism is an *intrinsically* incoherent conception.

Nevertheless, throughout history there has been no known alternative to encodingism — and there still isn't in standard approaches to representational phenomena — so the incoherence of encodingism, in its various guises, has seemed ultimately unsolvable and undissolvable, and therefore better avoided than confronted. The question "What else is there besides encodings?" still makes apparent good sense. Later we will outline an alternative that escapes the encodingism incoherence, but the primary focus in this book is on the consequences of the encodingism assumption. It does not attempt more than an adumbration of the solutions, which are developed elsewhere.

The Necessity of an Interpreter

The preceding discussion focused on the necessity of a *provider* of representational contents for the constitution of encodings, and on the impossibility of such a provider within encodingism itself. Here we will point out that there is a dual to this necessity of a provider that also has played a role in some contemporary work, and that is the necessity of an *interpreter*. Once an encoding representational content carrier has been created, an interpreter is required in order for that encoding to be used (for example, Gibson, 1966, 1977, 1979; see Bickhard & Richie, 1983; Shanon, 1993). Encodings in the formal symbol sense can be manipulated and generated with great complexity without regard to the representational content that they are taken as carrying, but if those resultant encodings are to be of any epistemic function, their representational content must be cashed in somehow. Encodingism (thus Artificial Intelligence and Cognitive Science) can neither explicate the function of the representational content provider, nor that of the representational content interpreter.

For computers, the user or designer is the provider and interpreter of representational content. This is no more problematic for the user or designer than is the interpretation of printed words or a picture as having representational content. As an attempt to account for mental processes in the brain, however, simply moving such interpretation accounts into the brain via analogy leaves unsatisfied and unsatisfiable the desire for a model of the user or designer per se — a model of the provider and interpreter of representational content. These functions are left to an unacknowledged and unexamined homunculus, but it is these unexamined intentional functions of the homunculus that are precisely what were to be modeled and understood in the first place. Such undischarged intentional homunculi in accounts of intentional phenomena are circular — they are aspects, in fact, of the basic circular incoherence of encodingism.

Most fundamentally, encodingism does not even *address* the *fundamental* problem of representation: The nature and emergence and function of representational content. Encodingism is intrinsically restricted to issues of manipulation and transformation of already-constituted *carriers* of representational content — carriers for some interpretive, intentional agent. That is, encodingism is not really a theory of representation at all: at best, it constitutes part of *one* approach to representational computations.

3

Consequences of Encodingism

LOGICAL CONSEQUENCES

Encodingist assumptions and presuppositions have many logical consequences. A large portion of these consequences are due to vulnerabilities of the basic encodingist assumptions to various questions, problems, objections, and limitations — and the ensuing attempts to solve or avoid these problems. We will survey a number of these consequent problems, and argue that they cannot be solved within the encodingist framework. We will analyze consequences of encodingism either in general conceptual terms, or in terms of distortions and failures of specific projects and approaches within Artificial Intelligence and Cognitive Science.

We begin with some classical philosophical problems that, we argue, are aspects of encodingist conceptions of or presuppositions concerning representation. Insofar as this argument is correct, then Artificial Intelligence and Cognitive Science face these problems as well by virtue of their presupposition of the general encodingist framework. In fact, we find manifestations of several of these classic problems in contemporary approaches.

Skepticism

There is more than one perspective on the basic incoherence of encodingism, and, in one or another of these perspectives, the problem has been known for millennia. Perhaps the oldest form in which it has been recognized is that of the argument of classical skepticism: If representational contents are carried or constituted only by encodings, then how can we ever check the accuracy of our representations? To check their accuracy would require that we have some epistemic access to the world that is being represented against which we can then compare

our encodings, but, by the encodingism assumption, the only epistemic access to the world that we have is through those encodings themselves. Thus, any attempt to check them is circularly impotent — the encodings would be being checked against themselves.

Idealism

A despairing response to this skeptical version of the encoding incoherence has classically been to conclude that we don't in fact have any epistemic access to the world via our encodings. We are epistemically encapsulated in our encodings, and cannot escape them. In consequence, it becomes superfluous to even posit a world outside those encodings — our basic encoding representations *constitute* all there is of our world. This response has historically taken the form of individual solipsism, or conceptual or linguistic idealism (Bickhard, 1995). Idealism is just a version of solipsism in the sense that both are versions of the assumption that our world is *constituted* as the basic *representations* of that world. Such "solutions" also yield at best a coherence version of truth.

Circular Microgenesis

Another perspective on the incoherence problem is the genetic one. Skepticism arises from questions concerning *confirmation* of encodings; the genetic problem arises from questions concerning the *construction* of foundational encodings. Not only can we not check our representations against an independent epistemic access to the world, but we cannot construct them in the first place without such an independent epistemic access to the world. Without such independent access, we have no idea what to construct. One version of this is the argument against copy theories of representation: we cannot construct copies of the world without already knowing what the world is in order to be able to copy it (e.g., Piaget, 1970a).

Incoherence Again

The incoherence problem itself focuses not on how encoding representations can be checked, nor on which ones to construct, but rather on the more foundational problem of how *any* representational content can be provided for a foundational encoding, and, thus, on how any logically independent encoding could exist at all. The answer is simple: it can't:

- There is no way to specify what such an encoding is supposed to represent;
- There is no way to provide it with any representational content;
- Thus, there is no way for it to be constituted as an encoding representation at all.

Non-derivative, logically independent, foundational, encodings are impossible. To postulate their existence, either explicitly, or implicitly as a presupposition, is to take a logically incoherent position.

Emergence

The root problem of encodingism is that encodings are a means for changing the *form* of representation — defining "• • •" in terms of "**S**" changes the form, and allows new things to be done: "• • •" can be sent over a telegraph wire, while "**S**" cannot. This is unexceptionable in itself. It becomes problematic only when encodings are taken as the foundational form of representation.

Encodingism encounters all of its circularities and incoherences at this point because encodings can only transform, can only encode or recode, *representations that already exist.* Encodingism provides no way for representation to emerge out of any sort of non-representational ground. Encodings require that representations already be available in terms of which the encodings can be constructed.

To attempt or to presuppose an encodingism, then, is to commit the circularity of needing to have representation before you can get representation, and the incoherence of needing to know what is to be represented before you can know what is to be represented (Bickhard, 1991b, 1991c, 1993a, in press-b). A strict encodingism requires that encodings generate emergent representations, and that is impossible for encodings.

On the other hand, there is no question concerning the fact that representation exists, and, for that matter, that encodings exist. Representational emergence, therefore, has occurred. At some point or points in evolution — and perhaps repeatedly in learning and development — representation emerged and emerges out of non-representational phenomena. These earliest forms of representation could not be encodings, since encodings require that what they represent be already represented, and, therefore, encodingism cannot in principle account for this emergence. A strict encodingism, in fact, implies that emergence is impossible (Bickhard, 1991b, 1993a).

The Concept of Emergence. The notion of emergence invoked here is nothing mysterious (though it can be conceptually complex: Bickhard, 1993a; Horgan, 1993; O'Conner, 1994). It simply refers to the fact that some sorts of things once did not exist, and now they do. At some point, they must have come into existence. If something that is of a different sort from what has existed before (even what has existed before *locally*, though the basic point can be made at the level of the whole universe) comes into existence, then that sort, or an instance of that sort, has emerged. Such a notion applies to molecules, galaxies, solar systems, patterns in self organizing systems, life, consciousness, and representation, among myriads of others. None of them existed at the Big Bang and they all do now. They have all emerged.

In most of these cases, we have some understanding of how they emerged, or at least of how they could in principle emerge. Such models of emergence are part of the general project of naturalism — of understanding the world in natural terms. In many of these cases, the understanding of emergence required a shift from a basic substance model of the phenomena involved — e.g., life as vital fluid — to a process model — e.g., life as a form of open system process. Basic substances cannot emerge. The Greeks' earth, air, fire, and water could not themselves emerge, but had to be in existence from the beginning. Substance approaches make emergence impossible to model — the basic substances are simply among the primitives of the approach.

That something has emerged is not strongly explanatory. It is a minimal explanation in that it explains why that something is existing now. But explanations themselves require explanations, and the fact of emergence is often not itself easily explained. The details of the emergence of life, for example, are still an open question. Substance models, however, have the consequence that any substance emergence is simply impossible, and close off the exploration before it can begin. Emergence, then, is neither strongly explanatory, nor is it mysterious. Emergence is simply a fact for many sorts of phenomena that itself needs to be explained, but that cannot be explained within a substance approach.

Representation has emerged, undoubtedly, countless times since the origin of the universe, though once is enough for the basic point. Representation, however, is still standardly conceptualized in substance terms — in terms of basic representational atoms out of which all other representations are constructed. The origin of the atoms themselves is mysterious, and must remain so as long as they are treated as

fundamental, because there is no way for them to emerge. Encodingism is built exactly on such an assumption of basic representational atoms — correspondence atoms — out of which other representations are to be constructed. But encodingism cannot account for the origin of those atoms. Encodingism presupposes such atoms rather than explaining them — that is its basic circularity.

Strict encodingism, therefore, cannot be true. There must be some other sort of representation that is capable of emergence, and, therefore, is not subject to the incoherence and circularities of encodingism.

4

Responses to the Problems of Encodings

FALSE SOLUTIONS

There have been, and currently are, a number of attempted solutions to partial realizations of the difficulties with encodings. Most commonly, however, the full incoherence of encodingism is not understood. Instead, some partial or distorted problematic consequence of the incoherence of encodingism is noted, and some correspondingly partial or distorted solution is proposed.

Innatism

One common response derives from the recognition that it is impossible to create, within encodingism, an encoding with new representational content. At best, derivative encodings can be constructed that stand-in for new combinations of already present encodings. But this implies that an epistemic system is intrinsically limited to some basic set of encodings and the possible combinations thereof. That is, the combinatoric space defined by a set of basic encoding generators constitutes the entire possible representational world of an epistemic system. Because that basic generating set of independent encodings cannot be itself generated by any known model of learning, so the reasoning goes, it must be genetically innate; the basic set of encoding representations must have been constructed by evolution (Fodor, 1981b).

One further consequence is that no interesting epistemic development is possible in any epistemic system (including human beings) because everything is limited to that innately specified combinatoric space. Another is the likelihood that the basic space of potential representations that are possible for human beings is limited concerning the sorts of things it can and cannot represent, and, thus, that

human beings are genetically epistemically limited to certain fixed domains of knowledge and representation (Fodor, 1983). Because these are fairly direct consequences of encodingism, Artificial Intelligence and Cognitive Science are intrinsically committed to them. But recognition of these consequences seems to have been limited at best. On the other hand, cognitive developmental psychology has been strongly seduced by them (see Campbell & Bickhard, 1986, 1987; Bickhard, 1991c).

The flaw in the reasoning, of course, is that the problem with encodings is logical in nature — an incoherence, in fact — and cannot be solved by evolution any better than it can be solved by individual development. Conversely, if evolution *did* have some mechanism by which it could avoid the basic incoherence — if evolution could generate emergent representations — then individuals and societies could avail themselves of that same mechanism. The assumption that the problem *can* be pushed off onto evolution invalidates the whole argument that supposedly yields innatism in the first place (Bickhard, 1991c).

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? 1) We can't have actual representations until we know what is to be represented. 2) But to know what is to be represented awaits millennial physics. 3) But physics cannot even *begin* until we have some sort of representations of the world. 4) Hence, 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. It's an example of a fatal

problematic of encodingism elevated to a purported solution to the problem of how to investigate representational phenomena.

Direct Reference

Another response to the impossibility of providing representational content to basic encodings has been to postulate a form of representation that has no representational content other than that which it encodes. The meaning of such an encoding *is* the thing that it represents. There is no content *between* the encoding element and the represented. Such "direct encodings" are usually construed as some form of true or basic "names," and have been, in various versions, proposed by Russell (1985), the early Wittgenstein (1961), Kripke (1972), and others. Again, this is a fairly direct attempt to solve the incoherence problem, but it seems to have been limited in its adoption to philosophy, and has not been much developed in either Artificial Intelligence or in Cognitive Science (though an allusion to it can be found in Vera & Simon, 1993).

Direct reference clearly simply sidesteps the incoherence problem. No way is provided by which such names could come into being, nor how they could function — how an epistemic system could possibly create or operate with such contentless representations. How are the "things" — which purportedly constitute the content of the names — to be known as the contents of those names? A classic philosophical stance to this question has been that that is a problem for psychology and is of no concern to philosophy. But if direct reference poses a problem that is logically impossible for psychology to solve, then it is illegitimate for philosophy to postulate it. Philosophy can no more push its basic epistemic problems off onto psychology (Coffa, 1991) than can Artificial Intelligence or psychology push them off onto evolution.

External Observer Semantics

Another response to the incoherence of encodings, and one currently enjoying an increasing popularity, is to remove all basic issues of representation outside of the systems or models being constructed, and simply leave them to the observer or the user of the system to be filled in as required. The observer-user knows that certain of the inputs, and certain of the outputs, are in such-and-such a correspondence with certain things in the world, and are thus available to be taken by that observer-user as encodings of those things in the world. There is no reason to postulate the necessity of any actual representations inside the system at

all. As long as it yields outputs that can be used representationally by the observer-user, that is sufficient. It is not even necessary to postulate the existence inside the system of any elements that have *any* particular correspondence to anything outside the system. And it is certainly not necessary to consider the possibility of elements inside the system that have the *known* such correspondences that would constitute them as encodings (again, for the observer-user to whom those correspondences were known) (Kosslyn & Hatfield, 1984).

This stance, however, does not solve any of the problems of representation, it simply avoids them. Pushing the representational issue outside of the system makes phenomena such as the generation of representational content, and intensional stances with regard to representational content, impossible to even address. It explicitly passes them to the observer-user, but provides no model of how any epistemic observer-user could possibly make good on the problem that has been passed to it. Among other consequences, this renders such an approach helpless in the face of any of the fundamental representational problems of observer-users. If we want to understand observers themselves, we cannot validly do so only by adversion to still further observers.

Internal Observer Semantics

The more "traditional" solution to the problem of representation within Artificial Intelligence and Cognitive Science has been to postulate not only representational correspondences for the inputs and the outputs of the system, but also for various elements internal to the system itself. Elements internal to the system are taken to be encodings that are manipulated and transformed by the system's operations.

Insofar as the encoding status of these elements is taken to be unproblematic, this is simply naive. Insofar as these elements are taken to be encodings by virtue of their being in factual correspondences with what they represent — the most common stance — it simply ignores the issue of how those correspondences are known or represented, and, in particular, how what those correspondences are *with* are known and represented. However *factual* such correspondences may be, the *representation* of such correspondences occurs only for the designer or observer or user, and, therefore, the internal elements (as well as the inputs and outputs) constitute encodings only for those designer-observer-users, not for the system itself (e.g., Newell, 1980a; Nilsson, 1991).

Factual correspondences do *not* intrinsically constitute epistemic, representational correspondences.

Allowing correspondences between internal states and the world may allow for the *simulation* of certain intensional properties and processes (those that do in fact involve explicit encoded representational elements in real epistemic systems — though there is reason to question how commonly this actually occurs), but ultimately the representational contents are provided from outside the model or system. Neither the external nor the internal observer-semantics view provides *any* approach to the foundational emergence or provision of representational content.

Some version of an observer semantics, whether external or internal, is in fact the correct characterization of the representational semantics of programs and their symbols. All such semantics are derivative and secondary from that of some already intentional, already representational observer — designer, user, or whatever. This is a perfectly acceptable and useful stance for design, use, and so on. But it is a fatal stance for any genuine explication or explanation of genuine representation — such as that of the observer him- or herself — and is impossible for actually trying to understand or construct intentional, representational, systems.

Observer Idealism

Standard approaches to the problem of representational contents typically either ignore it or hide it. In contrast, there is a radical approach that focuses explicitly on the observer dependence of encodings. Here, dependence on the observer-user for representational content becomes the purported solution to the problem — the only solution there is. Representational relationships and representational contents are *only* in the "eye" or mind of the observer or user. They are constituted by the observer-user taking elements in appropriate ways, and have no other constitution (e.g., Maturana & Varela, 1980).

Unfortunately, this approach simply enshrines an observer idealism. Such an observer is precisely what we would ultimately want Artificial Intelligence and Cognitive Science to account for, and such an observer idealism is in effect simply an abandonment of the problem — representation only exists for observers or users, but observers and users themselves remain forever and intrinsically unknowable and mysterious. Construing that observer as an intrinsically language-using observer

(Maturana & Varela, 1987) does not change the basic point: at best it segues from an individual observer idealism to a linguistic idealism.

Simulation Observer Idealism

A superficially less radical approach to the problem in fact amounts to the same thing, without being quite as straightforward about it. Suppose that, as a surrogate for an observer, we postulate a space of representational relationships — say, inference relationships among propositions — of such vast extent that, except for the basic input (and output) connections with the world, that structure of relationships itself constitutes "representationality," and, furthermore, constitutes the carrying of representational content. Then suppose we postulate: 1) a system of causally connected processes for which the network of *causal* relationships exactly matches the network of *representational* (propositional) relationships, and 2) that this system is such that the causal input and output relationships exactly match the epistemic input and output relationships. Finally, we propose that it is precisely such a match of causal with epistemic relationships that constitutes representation in the first place (e.g., Fodor, 1975, 1983; Pylyshyn, 1984).

Unfortunately, this approach simply defines representation in terms of matching relationships between causal phenomena and logically prior representational phenomena. As an explication of representation, this is circular. There is no model or explication of representational phenomena here — they are presupposed as that-which-is-to-be-corresponded-to, hence they are not addressed. The approach is at best one of simulation, not of explication.

The sense of this proposal seems to be that sufficient causal simulation will *constitute* instantiation, but the conceptual problem here is that the representational phenomena and properties to be simulated must be provided before the simulation/instantiation can begin. Representation is constituted by a causal match with representation, but there is no model of the representational phenomena and relationships that are to *be* matched. Those representational phenomena and properties are, of course, provided implicitly by the observer-user, and we discover again an observer idealism, just partially hidden in the surrogate of representational (propositional) relationships.

SEDUCTIONS

Transduction

Another important and commonly attempted solution to the problem of representational content is that of transduction. This is perhaps the most frequently invoked and most intuitively appealing — seductive — "solution," but it fares no better. Transduction is technically a transformation of forms of energy, and has no epistemic meaning at all. As used in regard to representational issues, however, it is taken as the foundational process by which encodings acquire representational contents.

The basic idea is that system transducers — such as sensory receptors — receive energy from the environment that is in causal correspondence with things of importance in that environment. They then "transduce" that energy into internal encodings of those things of importance in the environment. At the lowest level of transduction, these fresh encodings may be of relatively limited and proximal things or events, such as of light stimulations of a retina, but, after proper processing, they may serve as the foundation for the generation of higher order and more important derivative encodings, such as of surfaces and edges and tables and chairs (e.g., Fodor & Pylyshyn, 1981). In apparent support for this notion of transduction, it might even be pointed out that such transduction encoding is "known" to occur in the neural line (axon) and frequency encoding of the sensory inputs, and is "easily" constructed in designed systems that need, for example, encodings of temperature, pressure, velocity, direction, time, and so on.

What is overlooked in such an approach is that the only thing an energy transduction produces is a *causal* correspondence with impinging energy — it does not produce any *epistemic* correspondence at all. Transduction may produce correspondences, but it does not produce any knowledge on the part of the agent of the existence of such correspondences, nor of what the correspondences are with. Transduction may be functionally useful, but it cannot be representationally constitutive. Again, it is the observer or user who knows of that discovered or designed transductive correspondence, and can therefore *use* the generated elements, or *consider* the generated elements, as encodings of whatever they are in correspondence with (Bickhard, 1992a, 1993a).

Correspondence as Encoding:
Confusing Factual and Epistemic Correspondence

We consider here the most common error yielding naive encodingism: that discovered or designed factual correspondences (they do not have to be causal, e.g., Dretske, 1981) intrinsically constitute encodings. This error overlooks the fact that it is the observer or user who knows that correspondence, and therefore knows what the correspondence is with,[2] and therefore can construct the encoding relationship. The transduction model is simply a special case of this general confusion and conflation between factual correspondence and representation.

There is no explanation or explication in the correspondence approaches of how the system itself could possibly have any representational knowledge of what those correspondences are with, or even of the fact that there *are* any such correspondences — of how the system avoids solipsism. There is no explanation or explication of how the "elements that are in correspondence" — e.g., products of transductions — could constitute encodings *for the system*, not just for the observer-user (see Bickhard, 1992a, 1993a; Bickhard & Richie, 1983, for discussions of these and related issues).

That is, however much it may be that some changes internal to the system do, in fact, track or reflect external changes (thus maintaining some sort of correspondence(s) with the world), how the system is supposed to know anything about this is left unanalyzed and mysterious. Factual correspondences and factual covariations — such as from tracking — can provide information about what is being covaried with, but this notion of information is purely one of the factual covariation involved. It is a mathematical notion of "being in correlation with."

To attempt to render such factual information relationships as representational relationships, however (e.g., Hanson, 1990), simply *is* the problem of encodingism. Elements in covariational or informational factual relationships do not announce that fact, nor do they announce what is on the other end of the covariational or informational correspondences. Any attempt to move to a representational relationship, therefore, encounters all the familiar circularities of having to presuppose *knowledge* of the factual relationship, and *content* for whatever it is on the

[2] — and therefore has a bearer of the representational content for what the correspondence is with, and therefore can use that bearer to *provide* that content to the internal element-in-factual-correspondence —

other end of that relationship, in order to account for any representational relationship at all. Furthermore, not all representational contents are in even a factual information relationship with what they represent, such as universals, hypotheticals, fictions, and so on (Fodor, 1990b). Information is not content; covariation is not content; transduction is not content; correspondence is not content. An element **X** being in some sort of informational or covariational or transduction or correspondence relationship with **Q** might be *one* condition under which it would be useful to a system for **X** to carry representational content of or about **Q**, but those relationships do not constitute and do not provide that content. Content has to be of some different nature, and to come from somewhere else.

5

Current Criticisms of AI and Cognitive Science

The troubles with encodingism have not gone unnoticed in the literature, though, as mentioned earlier, seldom is the full scope of these problems realized. Innatism, direct names, and observer idealism in its various forms are some of the inadequate attempts to solve the basic incoherence. They have in common the presupposition that the problem is in fact *capable* of solution — they have in common, therefore, a basic failure to realize the full depth and scope of the problem. There are also, however, criticisms in the literature that at least purport to be "in principle" — that, if true, would not be solvable. Most commonly these critiques are partially correct insights into one or more of the consequences of the encodingism incoherence, but lack a full sense of that incoherence. When they offer an alternative to escape the difficulty, that "alternative" itself generally constitutes some other incarnation of encodingism.

AN APORIA

Empty Symbols

One recognition of something wrong is known as "the empty symbol problem" (Block, 1980; see Bickhard & Richie, 1983). There are various versions of this critique, but they have in common a recognition that contemporary Artificial Intelligence and Cognitive Science do not have any way of explicating any representational content for the "symbols" in their models, and that there may not *be* any way — that the symbols are intrinsically empty of representational content. There is perplexity and disagreement about whether this symbol emptiness can be solved by some new approach, or if it is an intrinsic limitation on our knowledge, or if the only valid stance regarding its ultimate solvability is

simply agnosticism. In any case, it is a partial recognition of the impossibility of an ultimate or foundational representational content provider within encodingism.

ENCOUNTERS WITH THE ISSUES

Searle

The Chinese Room. Searle's Chinese room problem is another form of critique based on the fact that formal processes on formal (empty) symbols cannot solve the problem of representation (Searle, 1981) — cannot "fill" those empty symbols with content. The basic idea is that Searle, or anyone else, could instantiate a system of rules operating on "empty" Chinese characters that captured a full and correct set of relationships between the characters input to the system and those output from the system without it being the case that Searle, or "Searle-plus-rules," thereby understood Chinese. In other words, the room containing Searle-executing-all-these-rules would receive Chinese characters and would emit Chinese characters in such a way that, to an external native speaker of Chinese, it would appear that someone inside knew Chinese, yet there would be no such "understanding" or "understander" involved.

The critique is essentially valid. It is a phenomenological version of the empty symbol problem: no system of rules will ever constitute representational content for the formal, empty symbols upon which they operate. Searle's diagnosis of the problem, however, and, correspondingly, his rather vague "solutions," miss the incoherence of encodingism entirely and focus on some alleged vague and mysterious epistemic properties of brains.

The diagnosis that we offer for the Chinese room problem is in three basic parts: First, as mentioned, formal rules cannot provide formal symbols with representational content. Second, language is intrinsically *not* a matter of input to output processing — see below — thus, no set of input-to-output rules adequate to language is possible. And third, genuine representational semantics, as involved with language or for any other intentional phenomena — as we argue below — requires the capability for competent interactions with the world. This, in turn, requires, among other things, skillful *timing* of those interactions. Searle reading, interpreting, and honoring a list of formal input-output rules provides no principled way to address such issues of timing.

The robot reply to Searle emphasizes the necessity for *inter*action between an epistemic system and its world, not just input to output

sequences. That is, the claim is that Searle's Chinese room misses this critical aspect of interaction (Searle, 1981). Our position would agree with this point, but hold that it is not sufficient — among other concerns, the timing issue per se is still not addressed.

In Searle's reply to the robot point (Searle, 1981), for example, he simply postulates Searle in the head of an interacting robot. But this is still just Searle reading, interpreting, and honoring various input to output rules defined on otherwise meaningless input and output symbols. The claim is that, although there is now interaction, there is still no intentionality or representationality, except perhaps Searle's understanding of the rules per se. Note that there is also still no timing.

Simulation? Our point is here partially convergent with another reply to Searle. Searle accuses strong Artificial Intelligence of at best *simulating* intentionality — the reply to Searle accuses Searle's Chinese room, whether in the robot version or otherwise, of at best *simulating* computation (Hayes, Harnad, Perlis, & Block, 1992; Hayes & Ford, in preparation). The focus of this point is that Searle is reading, interpreting, and deciding to honor the rules, while genuine computation, as in a computer, involves *causal* relationships among successive states, and between processing and the machine states that constitute the program. A computer running one program is a *causally* different machine from the same computer running a different program, and both are causally different from the computer with no program (Hayes, Ford, & Adams-Webber, 1992).

Searle's relationship to the rules is not causal, but interpretive. In effect, Searle has been seduced by the talk of a computer "interpreting" the "commands" of a program, so that he thinks that *Searle* interpreting such commands would be doing the same thing that a computer is doing. If a computer were genuinely interpreting commands in this sense, however, then the goal of intentional cognition would be realized in even the simplest computer "interpreting" the simplest program. A program reconfigures causal relationships in a computer; it does not provide commands or statements to be interpreted. Conversely, Searle *does* interpret such commands. He is at best simulating the causal processes in a computer.

Timing. In this reply to Searle, however, what is special about such causality for mind or intentionality or representation is not clear. We suggest that it is not the causality per se that is at issue — control relationships, for example, could suffice — but that there is no way of

addressing timing issues for interactions within the processes of Searle's interpreting activities. Furthermore, we argue below that this deficiency with regard to timing is *shared* by theories of formal computation, and thus, in this sense we end up agreeing with Searle again. In general, we accept Searle's rooms and robots as counterexamples to formal computational approaches to intentionality, but do not agree with either Searle's or other available diagnoses of the problem.

Interactive Competence. Note that Searle in the robot, or the room, could in principle be in a position to try to learn how to *reproduce* certain input symbols. More generally, he could try to learn how to control his inputs, or the course of his input-output interactions, even if they would still be meaningless inputs and outputs per se. If he were to learn any such interactive competencies, we claim he would in fact have learned *something*. Exactly what he would have learned, and especially how it relates to issues of representation, is not obvious. And, further, to reiterate, there would still be no timing considerations in any such interactions by Searle in his box. Nevertheless, we hold that something like this sort of interactive learning, especially when adequate interactive timing is involved, is the core of genuine representation and intentionality.[3]

Searle on the Mind. More recently, Searle (1992) has presented a major attack on cognitivism in a broad sense. Searle takes a number of positions and develops several arguments with which we are in agreement. He points out that, so long as syntax and computation are matters of ascription by an intentional agent, rather than being intrinsic, then any accounts of intentionality in terms of syntax or computation commit the homunculus fallacy — i.e., they account for intentionality with recourse to an intentional (homuncular) agent. He argues at length that syntax and computation are and must be such matters of ascription. Furthermore, Searle's discussion of his notion of the Background, and the sense in which it is necessary to more explicit intentionality, has intriguing resemblances to the implicit representationality of interactive skill intentionality (see the discussion of Dreyfus below). Our discussion does not proceed with the focus on consciousness that Searle advocates, but, nevertheless, there are several convergences.

On the other hand, Searle also takes a number of positions that we find troublesome. He endorses connectionism as somehow avoiding the

[3] Note the parallel with neural inputs and outputs — they too are meaningless per se, but does not preclude the interactions that they participate in from being meaningful.

problems that he attributes to cognitivism, missing the point that connectionist "representations" are just as much subject to the homunculus problem as those of standard cognitivism (for a less sanguine evaluation of connectionism by Searle, see his comments in Searle, 1990; Harnad, 1993a). He claims that functions are intrinsically ascriptive properties, and have no observer independent reality — sliding over the contribution that functions internal to a system can make to the very existence of the system itself, independent of any observer of that system (Bickhard, 1993a). And he continues to rely on mysterious, or at least undeveloped, notions of the brain "causing" consciousness. His analogy with water molecules causing the liquidity of the water is not a clarification: is this supposed to be efficient causality? If so, how? If not, then just what is Searle trying to say? Overall, we find ourselves in agreement with much of the general spirit of Searle's attack on cognitivism, but not at all in agreement with many of the specific arguments that he makes and positions that he takes.

The Cartesian Gulf. A major error that *seems* to underlie Searle's discussion is a rarely noticed relic of Cartesianism. It is not so much the assumption or presupposition that consciousness is a substance, but, rather, the assumption or presupposition that there is one singular gulf between the mental and the non-mental. Most commonly, this appears in the form of assuming that all mental properties must occur together: that a system that has one mental property must have them all. In contrast, we suggest (Bickhard, 1992c; see also the discussion of the evolutionary foundations of interactivism below) that there are many properties and processes of mentality, and that they have evolved in succession rather than having come into existence all at once at some unknown point in evolution. If so, then these multiple aspects of mentality will not form an indifferentiable unity. They will not be completely independent, since some will arguably require others to already exist — for their own existence or their own emergence — but mentality will form a perhaps multi-stranded evolutionary hierarchy of properties and processes rather than a single conceptual and evolutionary saltation.

The absence of any attempt on Searle's part to define consciousness is, on the one hand, understandable, but, on the other hand, provides a spacious hiding place for presuppositions such as the one that mentality is itself intrinsically unitary, with "consciousness" at its essential core. Searle's acknowledgement that it is not clear how far

down the evolutionary hierarchy consciousness might be found to extend appears to be one manifestation of this presupposition and of the sorts of perplexities that it can yield. What if some organisms exhibit perception or memory, but not consciousness? Is it possible for learning or emotions to occur without consciousness? The unitariness of Searle's undefinedness of consciousness makes such questions difficult to pose and to address.

Gibson

Gibson's critiques of standard approaches to perception have explicitly presented encodingism's necessity for an *interpreter* of representational content, and the necessity for a *provider* of representational content is implicit in another of his arguments (Bickhard & Richie, 1983). Gibson does not, however, develop the connection between these problems and encodingism per se. Gibson's critical stance, in fact, was overstated in such a way as to commit him to a version of encodingism — "direct" perception — in spite of his genuine and important partial insights into an alternative to encodingism (Bickhard & Richie, 1983). It should be noted that encodingism's need for an interpreter has nothing to do with whether such interpretation is or is not conscious. Gibson sometimes sounds as if that is what he is concerned with, and that is often how he is interpreted by critics (e.g., Ullman, 1980; Manfredi, 1986). Nonetheless, the basic issue is epistemic — the stand-in or carrier relationship must be *interpreted* in order for the representational content to function as such, whether or not such interpretation is conscious.

In spite of such problems, Gibson has provided the core of a non-encoding approach to perception (Bickhard & Richie, 1983). This is a major advance, especially since presumed sensory — perceptual — transduction is one of the strongest domains of encoding intuitions and models.

Piaget

Throughout his career, Jean Piaget argued against simple encoding models of knowledge. He explicitly presented the genetic argument against copy theories (e.g., Piaget, 1970a). His reliance on structuralism and, later, on information processing approaches, however, carried their own commitments to encodingism deep into his own epistemology

(Bickhard, 1988a; Campbell & Bickhard, 1986; Bickhard & Campbell, 1989).

Piaget's encodings, however, contained two rare and critically important insights. First, he recognized that representation must be grounded in and emergent from *action*. Second, he recognized that the most important form of knowledge was knowledge of *potentialities* — knowledge of potential actions, of the organization of potential transformations of environmental states, in Piaget's view — rather than passive knowledge of environmental *actualities*. These insights, along with Piaget's strong arguments for the necessity of the active construction of representations rather than their passive "impression" from the environment, moved Piaget far from a simple encodingism, but he was nevertheless unable to fully escape it.

Piaget's model of perception, for example, involves straightforward sensory encodings, while his notion of representational scheme involves structurally isomorphic correspondences with what is being represented (Bickhard, 1988a; Campbell & Bickhard, 1986; Bickhard & Campbell, 1989; M. Chapman, 1988). Piaget's argument against copy theories (Piaget, 1970a) points out that we would have to already know what we were copying in order to construct a copy of it — a circularity — so no notion of copying can solve the representational problem. But he then argues for representation as structural isomorphism with what is represented — something that sounds very much like a copy. Piaget's focus here was not on the nature of representation, but, rather on the nature of representational construction. Copying — passive impression from what is to be represented — does not work. Instead, representation must be constructed. But what is *constructed*, rather than copied, is still an isomorphic structure of correspondences — a copy.

We do not accept Piaget's basic notions of representation, but his constructivism is an essential part of understanding how the world is represented. If the ontogenetic or phylogenetic development from the most primitive representation — those of infants or primitive animals — to the most complex human adult representation cannot be understood within some purported model of representation, then no part of that purported model of representation is secure. Any model that cannot in principle account for such evolution and development of adult representation cannot be correct. Piaget's constructivism provides the skeleton for understanding that development (Piaget, 1954; Campbell & Bickhard, 1986; Bickhard & Campbell, 1989). Artificial intelligence and

Cognitive Science are still learning one of the fundamental lessons that he taught: it does not suffice to take *adult* representations as theoretical *primitives* of representation.

Maturana and Varela

Maturana and Varela have constructed a model of cognition and language in which they have, with great ingenuity, avoided both the transduction and the simulation-as-instantiation stances (Maturana & Varela, 1980, 1987). Unfortunately, as mentioned above, they have done so by constructing a pure and explicit observer idealism. For example, they correctly do *not* construe activities of the organism that are in factual correspondence with entities or events or properties of the environment as organism encodings for those entities or events or properties, but, instead, correctly place the recognition of those factual correspondences in an observer. They then, however, invalidly conclude that the representational relationship is constituted *only* by the distinctions that are made by such an observer. As with any observer idealism, this merely pushes all the basic epistemological issues into the unanalyzed and unanalyzable mysteries of the observer.

Dreyfus

Dreyfus (1979, 1981; Dreyfus & Dreyfus, 1986) has been a persistent critic of Artificial Intelligence aspirations and claims. The programmatic goals are impossible in principle, in his view, because the programme is based on fundamental misconceptions of the nature of understanding and language. In particular, the presuppositions of explicit, atomized, and context independent representations that are inherent in encodingism are deeply misguided and pernicious. Dreyfus does not develop his critique as a general critique of encodingism per se, although there are convergences, but instead brings to bear a hermeneutic perspective derived primarily from Heidegger (1962; Dreyfus, 1991; Guignon, 1983).

Atomic Features. A major focus of Dreyfus' critique is a presupposition of information processing approaches: that the world contains context independent atomic features — features that can be context-independently encoded. The problem is that the world does not (Dreyfus, 1991; Dreyfus & Haugeland, 1978). We would agree that it doesn't, and that this is one more reason why encodingism is untenable, but we would also argue that the fundamental flaws of encodingism

would remain *even if such atomic features **did** exist in the world*. In particular, the incoherence problem, among others, would not be altered by the assumption of such features. Factual correspondences with atomic features would still not constitute representations of them even if context independent atomic features did exist.

Skill Intentionality. Dreyfus' notion of skill intentionality, however, has a strong convergence with the interactive position that we are proposing as an alternative to encodingism (Dreyfus, 1967, 1982, 1991; Dreyfus & Dreyfus, 1986, 1987). The basic notion is that the intentionality of skills, which is usually taken as derivative from and subsidiary to standard representational intentionalities, whether mental or linguistic, should instead be taken as the more fundamental form of intentionality, out of which, and on the foundation of which, other forms of intentionality — such as representations — are constructed. Interactivism, in part, involves a convergence with that programmatic idea (see, for example, Bickhard, 1992c; Bickhard & Richie, 1983).

Criticizing AI. Dreyfus has been an outspoken voice in critiquing Artificial Intelligence assumptions and claims. And history, at least thus far, has borne out his criticisms over the dismissals of his opponents. The lesson of that history, nevertheless, has not yet been learned. We support most of the basic criticisms that Dreyfus has made and add some of our own. In fact, the encodingism critique yields its own critique of the typical representational atomism in Artificial Intelligence — and covers as well contemporary connectionist and analog proposals.

Connectionism. In contrast to Dreyfus (1992; Dreyfus & Dreyfus, 1988), then, we are not at all sanguine about the prospects for contemporary connectionist approaches. There is, in fact, something surprising about the major proponent of know-how and skill intentionality expressing such acceptance of an approach in which most models do not have *any* interaction with their environment at all, and, thus, cannot have *any* know-how or skill at all. (Even for those that do involve some form of environmental interaction, this is an engineering level add-on, and has no relevance to the basic theory of connectionist *representations*.) In the end, connectionist systems, like Good Old Fashioned AI systems, just passively process inputs. The modes of processing differ, but the arguments we present below show that that difference in mode does nothing to avoid the fundamental basic problem that afflicts both approaches equally.

Situated Cognition — Reinforcement Learning. Dreyfus (1992) also expresses interest in the situated cognition of Chapman and Agre, and in an approach called reinforcement learning. We share his judgment that these approaches involve major advances, but, nevertheless, they too commit the basic encodingism error. Reinforcement learning, for example, requires (among other things) a built-in utility function on the inputs — the system has to already know what inputs to seek, and usually must also have some loss function that is defined on errors. Such an approach can be practically useful in certain circumstances, but, as a general approach, it requires that critical and potentially complex knowledge be already built into the system before it can learn. That is, it requires already existing knowledge in order to learn knowledge. It is crucial to realize that this approach does not use these types of built-in knowledge just for convenience. Rather, the built-in knowledge is essential for the later learning of the system, and the model offers no account of how the initial knowledge can be learned. As a *general* approach, this immediately yields a vicious infinite regress — the regress of impossible emergence.

We claim to provide a model that does not fall to these problematics, and, in fact, *does* provide an approach to know-how and skill intentionality. In effect, we agree with Dreyfus about the necessity for some sort of holism in addressing human-level intentionality, but disagree about its ultimate importance. Holism without interaction, such as in connectionist systems, does not avoid the incoherence problem. Conversely, interactivism easily covers many "holistic" phenomena (see, for example, the discussions of the frame problems below, or of an interactive architecture), but a kind of holism is a consequence, not the core, of the interactive approach.

Hermeneutics

Historically, hermeneutics derives from the interpretation and understanding of historical texts; it emphasizes the intrinsic situatedness of all understanding, and the intrinsic linguistic and historical nature of all such situations of understanding (Bleicher, 1980; Gadamer, 1975, 1976; Howard, 1982; Warnke, 1987). Understanding is inextricably embedded in linguistic historical situations because understanding is always a matter of hermeneutic interpretation and reinterpretation — interpretation and reinterpretation, in turn, is always in terms of language, and is, therefore, intrinsically constituted within and from the social, cultural, and historical

sedimented ontology of that language. To try to eliminate that context-dependent embeddedness in language and history in favor of atomized, finite, context independent representations inevitably does radical violence to the ontologies involved.

Clearly there is a general convergence between the hermeneutic position and the encoding critique proposed here (and it is even stronger when the alternative to encodingism that we offer is considered), but there is also a danger which hermeneutics does not seem to have avoided. If understanding is ontologically a matter of interpretation, and interpretation is ontologically constituted in terms of historically situated language, then it is seductive to conclude that all understanding is linguistic in nature, and, therefore, that language provides and circumscribes our epistemology and our world. In other words, it is seductive to conclude that: "That which can be understood is language." (Gadamer, 1975, p. 432), or "Man's relation to the world is absolutely and fundamentally linguistic in nature." (Gadamer, 1975, p. 432), or " ... we start from the linguistic nature of understanding ... " (Gadamer, 1975, p. 433), or "All thinking is confined to language, as a limit as well as a possibility." (Gadamer, 1976, p. 127).

Unfortunately, such a position lifts all epistemology and ontology into the realm of language as an absolute limit — it constructs a linguistic idealism. But linguistic idealism is just a variant of observer idealism, it is a social-linguistic-idealism, a socially located solipsism. All issues of the non-language world — of the relationships, both epistemological and interactive, *between* the individual and that nonsocial, nonlanguage world; of the embodiment of the individual *in* that nonsocial, nonlanguage world; and, still further, all issues of the nature of the individual as being materially, developmentally, and epistemologically *prior* to the social linguistic world; and of the constitutive and epistemological *relationships of* such individuals *to* that social linguistic world — all such issues are either ignored, or are rendered as mere issues of interpretation and discourse *within* that social linguistic world (Bickhard, 1993b, 1995).

When specifically pressed, Gadamer, and, presumably, most other hermeneuticists, do not want to deny that non-hermeneutically constituted reality (e.g., Gadamer, 1975, p. 496), but there is no way within hermeneutics per se to acknowledge it, or to approach questions as to its nature or its relationships to the domain of hermeneutics. In other words, there is no way to consistently avoid a linguistic idealism. This can make it quite difficult to make use of the insights that are present in the

hermeneutic approach without either explicitly or implicitly committing to such a linguistic idealism (e.g., Winograd & Flores, 1986). (This linguistic idealism of hermeneutics is strongly convergent with the later Wittgenstein, who is, in fact, sometimes counted as a hermeneuticist [e.g., Howard, 1982], even though his historical roots differ from those of Heidegger and Gadamer. Wittgenstein's linguistic idealism, or at least the possibility of such, is discussed further in Bickhard, 1987.)

6

General Consequences of the Encodingism Impasse

REPRESENTATION

The incoherence of encodingism as an approach to the nature of representation has differing consequences for differing parts of Artificial Intelligence research and Cognitive Science. Most centrally, phenomena of perception, cognition, and language cannot be adequately understood or modeled from an encoding perspective. These form the backbone of cognition as classically understood. On the other hand, the incoherence of encodingism arises from the presupposition that encodings form the essence, or at least a logically independent form, of representation, and many research goals, especially practical ones within AI, do not necessarily depend on that programmatic supposition. Many practical tasks can be solved quite satisfactorily within a user dependent semantics for the "symbols" involved — for example, the word processor upon which this is being written. But *all* of the basic *programmatic* aspirations of the fields involve representation — essentially — and, therefore, none of those aspirations can be accomplished with current encodingist frameworks.

LEARNING

The encodingism presuppositions of explicit, atomized, and context independent representations are always potentially a problem, and become more of one the more the task depends on the real properties of representation, reasoning, understanding, and communication. One illustrative domain in which this appears with particular clarity is that of learning.

First, note that learning involves the construction of new representations, but, within encodingism, the only new representations possible are just new combinations of some original set of observer-user dependent encodings. That is, all possible representations-to-be-learned must be anticipated in the combinatoric space of the generating set of basic encodings. This anticipation must be done by the designer in the case of AI learning research, and by evolution in the case of human beings. In practice, anticipations in that space of combinations have tended to be quite shallow. But genuinely new representations are prohibited by the incoherence of new basic encodings, and, in the general case, reliance on observer-user dependent semantics for the construction of "new" encodings — new elements or atoms — merely abandons the task of genuine machine learning.

Even in the most sophisticated expert systems, the spaces of possible problem categorizations and of possible problem solutions are, at best, simple pre-designed combinatorial spaces, with the possible combinatorial constructions serving to model the problematic systems under investigation for purposes such as trouble-shooting, simulation, and so on (Clancey, 1992c). In simpler cases, the combinatorial space is flat, and the expert system heuristically classifies into *nominal* classes of predefined problem types with predefined solution types (Clancey, 1985). As enormously useful as these can be, they do not engage in the learning of new representational atomic units.

Within an encoding framework, for example, a repair robot would have to contain in its data structures a combinatoric space of representations that would be fully adequate to *all possible* breakdown situations it might encounter. If the repair robot, for example, had encoding atoms only for electrical phenomena, then, no matter how competent it might be for electrical phenomena, it would be at a loss if the plumbing leaked, or a support beam buckled, or a brick fell out, or ... just choose something outside of the given combinatoric space. This of course means that the programmer would have to at least implicitly anticipate the space of *all* such possible breakdowns.

Such omniscient anticipations are clearly impossible. The point of learning, after all, is to succeed when anticipations have failed. A repair robot dependent solely on encodings for its representations would be at a loss whenever it encountered a novel situation. This might not render it totally useless — it might even be extremely useful for most actually encountered situations in certain circumscribed domains — but it could

not engage in any true learning, and would likely be frequently helpless in any but the most closed conditions. The anticipation problem would be unboundedly open, for example, concerning the space of possible breakdowns on an unmanned space station.

A second sense in which encodingism makes genuine learning impossible turns on the fact that learning requires error, and genuine error cannot be defined in a strict input-processing encoding system. Error requires some standard, from the perspective of the system itself, that can be successfully satisfied or fail to be satisfied. Learning ultimately turns on how to avoid such error. Learning requires some sort of constructive variation of system organization so long as errors are encountered.

User or designer provided error criteria simply import from outside the system the necessary supplements for learning to occur. These are no more a general solution to the problem of learning than a user or designer semantics is a solution to the problem of representation. Learning with designer provided error criteria is also fixed, unless further user or designer interventions occur: such a system cannot learn new kinds of errors.

A system with *designer provided* error (goal) criteria and *designer provided* combinatoric data spaces could use feedback to select from within that combinatoric data space some combination that minimizes the defined error. Selection from pre-defined spaces of possibilities on the basis of feedback about pre-defined goals or error criteria is what is called learning within the framework of Machine Learning. Again, this might be very useful in some circumstances. But it cannot be a general approach to or solution to learning because it requires all of the prior knowledge of what counts as error and success, and what the anticipatory combinatoric space is that supposedly contains the solution, to be already provided to the system before the system can function at all. This approach, then, involves massive requirements of prior knowledge in order to get knowledge. It is really "just" the exploration of a predefined space for a satisfaction to a predefined criterion — at best, a very weak and limited form of learning. No learning of genuinely new error criteria, and no learning outside of the predefined combinatoric space, is possible.

The requirement for error criteria and error signals in order for learning to occur yields further problems for encoding approaches. We illustrate with three of them. The first is that a *strict* encoding system will simply encode in some way or another (or fail to encode) all inputs. Without something in addition to the processing of encoded inputs into

other encodings, there is no way to classify some inputs as simply inputs to be processed, and some as constituting success or error. The very distinction between a *feedback* input and "just another input to be encoded and processed" must itself be pre-built into the system. Inputs are inputs for an encoding system, and that is all there is. Learning requires error, and error requires criteria that encodingism per se cannot provide. Error is not just one more thing to be encoded.

A shift in perspective on this same point highlights our second point, an encounter of encodingism with skepticism in the context of learning. Learning in any general sense is in response to error, but if a system is a strict, passive, encoding system, then it has no way to check if its encodings are in error. If such a check is attempted, the system will simply re-encode in the same way — a way that is potentially errorful from an *observer* perspective. The system itself, however, has no way of distinguishing such "error." The system cannot check its encodings against what is supposed to be encoded; at best, it will simply "encode again." A pure encoding system is caught in a solipsistic epistemology, and, since solipsism provides no ground for error checking, a pure encoding system cannot learn.

The third problem involves feedback. Consider a machine learning system with as much built into it as possible — concerning error, concerning what counts as feedback, and concerning the generation of a combinatoric space of possibilities. Note that this system cannot be purely a passive encoding system: it requires interaction with some environment in order to *derive* feedback so that it can search in its combinatoric space. It is not a novel point that error feedback can be required for learning (e.g., Bickhard, 1973; D. Campbell, 1959, 1974; Drescher, 1991; Piaget, 1971, 1985; Popper, 1965, 1972), but the import of that requirement for interactive feedback *for the nature of representation itself* has not been understood. The basic intuition of that import, which we will elaborate later, is that the system ultimately learns what outputs to emit under what prior internal interactive conditions. It learns forward-looking anticipations of what actions and interactions would be appropriate, rather than backward looking analyses of the environmental causes of its current states. That is, it learns interactive knowledge. It does not learn correspondences between its inputs and its world. If representation can be learned, then representation must be somehow constituted in such interactive knowledge, not in input-to-world correspondences — not in encodings.

Encodingism impacts issues of learning, then, in at least three ways: 1) the space of all possibilities that can be searched must be predefined for the system, 2) error criteria and error signals must be predefined for the system, and 3) even with such predefinitions, the system cannot be just an information processor — it must generate *interactive outputs* in order to generate feedback. Learning, then, is one domain, though not the only one, in which the in-principle incoherence of encodingism manifests itself for even the most practical goals.

THE MENTAL

At the level of *programmatic aspirations*, however, the encodingism incoherence renders both Artificial Intelligence and contemporary Cognitive Science simply bankrupt (Bickhard, 1991b, 1991c, 1992c, 1993a). Encodingism *cannot* explicate or explain or model the phenomena of representation, nor any of the myriad other mental phenomena that involve representation — perception, memory, reasoning, language, learning, emotions, consciousness, the self, sociality, and so on. And any Artificial Intelligence or Cognitive Science model that *does* simulate or approximate in some way some such phenomenon will, by virtue of that encodingism, be a distorted and misguiding foundation for any deeper understanding, or for further extension of the model. Encodingism is a foundationally flawed approach to the domain of the mental.

WHY ENCODINGISM?

If encodingism is in such trouble, why is it so dominant — and why has it been so dominant for such a long time? What is the appeal of encodingism? There are several reasons for this appeal (Shanon, 1993).

The first is simply that external representations in general *are* encodings. Paintings, statues, maps, blueprints, ciphers, military codes, computer codes, and so on, form a vast realm of interpreted encoded representations, and it is only natural that these are the forms that are most readily taken as constituting representation. It is apparent that mental representations cannot be *identical* to any such external representations, but it is not so apparent how fundamentally different mental representation must be.

Related to this ubiquity of external representations is the point that these are all structural representations, either structures of objects or of

properties or of events. Objects and their properties are among the first cognitions available developmentally, and substances and their properties have universally been among the first sorts of ontology proposed for the subject matter of virtually all sciences. If the nature of representation is being explored, and an object or substance approach is assumed, then some sort of structural correspondence model — some sort of encodingism — is a natural starting place. Movement to a process model takes time, and requires more sophisticated notions of process. These notions typically develop, whether ontogenetically or historically, within the framework of, and therefore later than, prior object and substance approaches. Process models come later than object or substance models, naturally. Investigations of representation have "simply" not yet made the shift.

A third reason that encodingism has maintained such a grip on models of representation is that the problematics of encodingism form a vast and intricate maze of red herrings. There are myriads of versions; myriads of problems to explore; myriads of potential fixes for each one — and more versions, problems, and potential fixes are being discovered all the time. Encodingism frames one of the most complex programmes ever, yet it has not been at all apparent that it is a flawed programme, nor where and how deep that flaw might be even when some such flaw has been suspicioned. Many fixes that have purported to overturn the tradition turn out to be just another version of it (for an analysis of one contemporary example, see Bickhard, 1995).

It is not a mystery, then, that encodingism has been and remains the dominant programmatic approach to representation. Encodingism seems obvious in the many examples externally available. Developmentally, it is a necessary starting point. And it provides millennia worth of red herrings to follow and cul-de-sacs to explore.

II

INTERACTIVISM: AN ALTERNATIVE TO ENCODINGISM

7

The Interactive Model

Encodingism is such an intuitive position that it seems to be obviously true. Even when various problems with it are discovered, they are most easily assumed to be problems with particular models or formulations, not with the approach per se. New and better formulations within the same framework that overcome the deficiencies of current models is the promissory note that constitutes something as a programme rather than as being a model or theory itself. Artificial Intelligence and Cognitive Science are programmatic in precisely this sense. Unfortunately, this continuous reliance on the next as-yet-unformulated model or theory to remedy current deficiencies presupposes that the programme per se is in fact foundationally valid. No amount of construction of particular models will ever in itself uncover (much less fix) a foundational programmatic flaw — that requires in-principle arguments that are directed against the defining presuppositions of the programme. Otherwise, it is always easy to assume that the next theory, or the next decade, will provide the fix.

Such programmatic failures have been the fate of other scientific paradigms, such as behaviorism, associationism, and the two-layer Perceptron approach to pattern recognition. Even with in-principle arguments, however, it is easier to grasp the inadequacy of an approach when an alternative is available. A better solution helps in diagnosing and understanding the problems with a flawed solution. Conversely, it can be difficult to discern an in-principle difficulty, or to accept the validity of an in-principle argument, if there is no alternative to consider, and no alternative perspective from which to view the issues. If no alternative seems conceivable — What else is there besides encodings? — then the in-principle arguments against a presupposition may themselves be taken to be *their own reductios* by virtue of claiming that an "obvious," and obviously necessary, presupposition is false.

This has in fact been the fate of the skepticism-solipsism dilemma throughout history. Many attempts — all unsuccessful — have been made to disprove or dissolve skepticism. Contemporary approaches have generally either argued that it is self-contradictory in that it in some way presupposes the very world it purports to question, or that it is absurd in leading to a denial of what is epistemologically necessary — the existence of the world. There are many ingenious variants on these positions (Annas & Barnes, 1985; Burnyeat, 1983; Groarke, 1990; Popkin, 1979; Rescher, 1980; Stroud, 1984), but they all involve at root the presupposition that encodingism does in fact constitute the only approach to epistemology. To accept the skepticism-solipsism dilemma, or any of its variants, as themselves reductios of encodingism would yield a deep perplexity as long as no alternative is available.

There is an alternative, and it is in fact unlikely that the above critique of encodingism could have been discovered or understood in its present scope without the background and perspective of this alternative. The alternative is an alternative conception — an **interactive** conception — of the *nature* of representation, with consequences throughout epistemology and psychology. As such, it becomes understood only to the extent that its ramified consequences throughout philosophy and psychology have been explored. That is a massive — in fact, a programmatic — task that will not be attempted here. We do wish to present enough of this alternative, however, to at least indicate that it does exist, and to be able to make use of some of its parts and aspects in later discussions.

BASIC EPISTEMOLOGY

Representation as Function

Encodingism focuses on *elements* of representation. Interactivism requires a shift to a view of representation as being a ***functional*** *aspect of certain sorts of system processing*. This shift from representations as elements to representation as function is critical. It is possible, within this functional view, to set up systems of elements that serve differentiated and specialized representational functions, and to create encoding stand-ins for serving those functions. That is, it is possible to construct *derivative* encodings on an interactive functional representational base. But, from the interactive perspective, such encodings can only be defined on, and can only emerge from, such an already existing interactive representational base. Thus, it provides an account of the “ground” or

"foundation" for representational content that encodingism cannot. Furthermore, the properties of interactive derivative encodings are not identical to the presupposed properties of classical encodings (Bickhard & Richie, 1983).

Interactivism, then, provides a *functional* model of representation. That is, it presents a functional explication of representat***ion*** (or representing), 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 (Van Gulick, 1982). Conversely, anything that does function appropriately for a system will by virtue of that be a representation, or serve the function of representation, for that system. This view is in stark contrast to the encodingist conception of context-independent elements carrying representational content in virtue of being in some correspondence relationship.

This relatively simple — and incomplete — 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.

Representation as function has a broad convergence with notions of *meaning as use*, as in the later Wittgenstein or in some conceptions of programs. But, we maintain, representation must in addition involve some sense of "use" that can be wrong, and representation must capable of being wrong *for the system itself* (Bickhard, 1993a, in preparation-c). These criteria are not met, and are generally not even addressed, by contemporary approaches. The first point, we argue, requires action and

interaction, not just input and internal processing, while the second requires (normally) goal-directedness.

Similarly, the interactive model of representation as function is, strictly, a version of *wide functionalism* in the sense that the required functional relationships involve the environment as well as functional processes internal to the system. But, just as standard ways of elaborating functionalist models are infected with encodingism in their development beyond the basic intuitions (Bickhard, 1982, 1993a; Bickhard & Richie, 1983), so also are developments within wide functionalism (Block, 1986; Harman, 1982, 1987). Among other divergences, wide functionalist approaches in general do not recognize the fundamental necessity that the functional processes close on themselves, circularly — that they form *inter*actions. Correspondingly, they cannot address the criterion of representations potentially being wrong *for the 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 two additional logical necessities are required. These additional requirements are the possibilities of error and of error for the system. First, the functional differentiation of system activities must be in some sense epistemically related to some environment being represented. Second, those differentiations must 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.)

Abstract Machines. Just as the interactivist position has affinities and differences from standard notions of meaning as use and with wide functionalism, it also has affinities and differences with available formal mathematics for "use," for function and functional processes — the mathematics of abstract machines and abstract machine processes. This is in fact the formal mathematics that underlies classical Artificial Intelligence and Cognitive Science. An *automaton*, for example, is one simple version of such an abstract machine. Automata theory conceptualizes a machine as being in one of some set of possible abstract internal machine states, and as moving from state to state in state transitions that are triggered by the receipt of particular inputs into the system. A particular pair of current state plus current input, then,

determines the next state (Eilenburg, 1974; Ginzburg, 1968; Hopcroft & Ullman, 1979).

A simple *recognizer* is an automaton with some designated start state, and some set of designated final states. A string of inputs will trigger various transitions from internal state to internal state, leaving the automaton in some particular state when the input string ends. If that internal state at the end of the receipt of the input string is one of the designated final states, then that automaton is said to *recognize* that input string — the automaton distinguishes those strings that yield some designated final state(s) from those that do not (Eilenburg, 1974; Ginzburg, 1968; Hopcroft & Ullman, 1979). It distinguishes those strings by virtue of *in fact* ending up in one of the designated final states.

Note that any such final state is not an encoding for the automaton itself. If a final state is in any correspondence with anything, the automata doesn't know it. A final state can only be functional for the system itself by influencing further processing in the system — or in some broader system. As mentioned above, we will argue that some version of such influence on, of such control of, further processing is the locus for the emergence of genuine representational content. Our critical point here, however, is that such a final state is *not* an encoding for the automaton. Neither do the inputs to an automaton constitute representations for that system. Again, if there are any factual correspondences involved, the system does not know about them.

These points seem relatively obvious for automata. But exactly the same points hold for more complicated and computationally powerful machines, all the way to and including Turing machines and their programming languages. At this level, however, there is the overwhelming temptation to interpret the inputs and the internal states (and structures of states) as representational encodings — to interpret symbols, frames, arcs, nodes, pointers, and so on as representational encodings. Yet nothing in principle has changed in moving from automata to programming languages. The increased power involved is increased computational power, not representational power. It is, for example, increased power with respect to what classes of strings of input elements can be computationally "recognized" or differentiated, not increased power with respect to what the symbols and states can represent.

Nevertheless, we claim that there are some fruitful aspects of these abstract machine and abstract process conceptions. They can pick up, in

fact, on the intuitions of influence on later process, of meaning as use. We turn now to how to make good on those conceptions and intuitions. We note that some of the characteristics that will be crucial to the interactive model that are absent from such abstract machine notions are outputs, *inter*actions, goals, feedback, and timing of interactions.

Epistemic Contact: Interactive Differentiation and Implicit Definition

Consider a system or subsystem in interaction with an environment. The course of that interaction will depend in part upon the organization of the system itself, but in part it will also depend upon the environment being interacted with. Differing environments may yield differing flows of interaction. Correspondingly, differing environments may leave that (sub)system in differing final internal states or conditions when the interaction is "completed." Such possible internal final states, then, will serve to *differentiate* possible environments — they will differentiate those environments that yield internal final state **S13** from those that yield internal final state **S120**, and so on. A possible final state, correspondingly, will *implicitly define* the class of environments that would yield that state if in fact encountered in an interaction. These dual functions of environmental differentiation and implicit definition are the foundations of interactive representation.

Note, however, that a final state will not indicate anything at all about its implicitly defined environments — except that they would yield that final state. A possible final state will be in *factual* correspondence with one of its implicitly defined environments whenever that state is in fact reached as a final state, but the state per se contains no information about what that correspondence is with — the relationship to the corresponding class of environments is purely implicit. *Thus there is no semantic information, no representational content, available that could make that final state an encoding.* Note that this condition of being in a factual correspondence with unspecified environmental properties or conditions is precisely the condition of actual (sensory) transducers — only in the observer can there be the knowledge of both sides of the correspondence that allows the construction of an encoding.

In effect, such possible final states (or internal system indicators thereof) constitute a basic representational *function* without themselves bearing any representational *content* — nothing is represented *about* the implicitly defined class of environments except that it is different from the other differentiated classes. This seemingly small separation of *being*

a representation (a differentiator, in this case) from *bearing* representational content is a fundamental difference between interactivism and encodingism, and makes interactivism invulnerable to the fatal flaws of encodingism, including the incoherence problem and the skepticism-solipsism dilemma. In particular, an interactive differentiating final state does not require that what is being represented be already known in order for it to be represented. It is precisely that requirement for encodings that yields the incoherence of foundational encodings. Foundational encodings are supposed to provide our basic representational contents, yet they cannot be defined or come into being without those representational contents being *already* provided — an encoding is a representation precisely because it already has representational content.

Representational Content

Thus far, however, we have only indicated how something could serve an implicit representational function without specifying how it could have representational content. Representational content must be constituted somehow, and it remains to indicate how interactivism can account for that content without simply providing it from the observer-user as in the case of encodingism.

The basic idea is that other subsystems in the overall system can use the differentiations in those final states to differentiate their own internal goal-directed processing. For example, if subsystem **T94** is functioning with goal **G738**, and if subsystem **T77** has ended with final state **S13**, then **T94** should select strategy (interactive procedure; interactive system organization) **St3972**, while if **T77** ended with final state **S120**, then **T94** should select strategy **St20**. The final state that **T77** reaches serves to differentiate, to select, the activities of **T94**; final state **S13** indicates strategy **St3972**, and final state **S120** indicates strategy **St20**. In general there may be vast and complex organizations of such interactive processing selection dependencies.

The critical point to note is that such processing selection dependencies do constitute representational content about the differentiated environmental classes. In the above example, **S13** type environments are predicated to have interactive properties appropriate to strategy **St3972**, while **S120** type environments are predicated to have interactive properties appropriate to strategy **St20**. These representational contents are constituted in the possible selection-of-further-processing

uses that can be made of the differentiating final states. Conversely, the final states and their indicators indicate the further interactive properties appropriate to whatever selections of further interaction that might be made on the basis of those final states.

Final states that are in such further-processing selection relationships thereby indicate further interactive properties and potentialities of the implicitly defined environments. Furthermore, such indications of interactive potentialities can be *wrong*, and can be *discovered to be wrong* by the failure of the system to advance toward its goal — as in feedback systems, servomechanisms, and trial and error learning. Representational content can emerge, be added to, and be changed by changes in the organization of the overall system, particularly by changes in the selections made of possible further processing. The representational content comes *after* the existence of the implicitly defining, differentiating, representation, both logically and constructively. Representational content, in this view, is defined as indications of potential further interactions (Bickhard, 1992a, 1993a).

It is important to note that the conception of "goal" that is needed in this model does *not* require that goals be themselves representations. If goals did have to be representations, then representation would have been explicated in terms of representations (goals) — a vicious circularity. Instead, the goals in this model need only be internal functional switches that, for example, switch back into a trial and error interactive process or to a learning process under some conditions (functional failure), and switch to further processing in the system under other conditions (functional success) (Bickhard, 1993a). A goal of maintaining blood sugar level above some level, for example, need not involve a representation of blood sugar level; it requires only that some internal functional switching condition, with appropriate switching relationships, be *in fact* dependent on blood sugar level. Such functional goals *can* be based on subsidiary representational processes, but they do not *require* representation, and, therefore, do not defeat the modeling of representation out of non-representational organization. This stands in strong contrast to common conceptions of goals as *representing* environmental goal conditions.

Representation without goals. There is, in fact, a more primitive version of interactive representation that does not require goals at all. In this version, the indications are of possible interactions and of the ensuing possible internal outcomes of those interactions (Bickhard, in preparation-

c). Such indications might be useful, for example, in selecting which among the possible interactions at a given time is to be executed — selection based on the indicated subsequent internal outcomes. Whether or not those indicated outcomes are in fact reached is a system detectable condition — a purely functionally detectable condition — and failure to reach indicated conditions falsifies the indications. It is important to note that this potentiality for error in the indications is error for the system, of the system, and detectable by the system. In particular, this is not just error imputable or diagnosable by some external observer of the system. Such indications, then, have truth value for the system. Such indications are system representations, without goals.

Functional goals. On the other hand, a system, especially a living system, is not going to actually detect such error in its indications unless it can do something with that information — information that the indicated conditions do not exist. What could it do with such error information? It could reiterate the interaction, try a different interaction, or invoke some learning procedure. In any such case, we have criteria for continuing to pursue the condition and criteria for exiting on to other processes: we have functional (though not necessarily representational) goals.

The logical function that goals serve in the interactive model is to provide criteria for error. We have just shown that there is a more primitive manner in which error could be detectable in and for a system, but that real systems are likely to actually generate error information only if they can do something with that information. What they do with error information is to try various possibilities for eliminating or avoiding such error, which constitutes a functional goal.[4] In real interactive systems, then, error information, thus representation, will generally involve functional goals, and we will continue to characterize interactive

[4] In complex systems, error information may influence the course of *internal* interactive processes among multiple internal parallel subsystems. Goal-directedness can be an emergent phenomenon of such internal interactions (Steels, 1991, 1994; Maes, 1994; Beer, 1990; Brooks, 1991a; Cherian & Troxell, 1994a, 1994b, in press). Such emergence affords important architectural possibilities, but the complexities of the analyses involved in modeling or designing such systems (e.g., functional *themes*: Bickhard & Richie, 1983; Bickhard, 1992c) are only indirectly relevant to our basic point that a system will generate error information only if it can do something with that information. Whatever it does with that information that constitutes the detected internal condition as a *functional* error condition will also emergently constitute it as a *representational* error condition. Therefore, as mentioned, we will continue to characterize interactive representation as requiring goals.

representation as requiring such goals. Goals too, however, have simpler and more complex examples.

Some examples. A bacterium, for example, might differentiate its world into two categories: swim and tumble (D. Campbell, 1974, 1990). An external observer can note that "swim situations" seem to be those in which things are getting better — either by virtue of swimming up a food gradient, for example, or by swimming down a gradient of noxiousness — while tumble situations are those in which things have been getting worse — down a food gradient or up a noxiousness gradient. Transduction encodingism would suggest that the chemical transducers in the bacterium encode (the first time-derivative of) various foods and noxious stuff, from which the bacterium would then have to infer the proper action. From the interactive perspective, however, the fact that the transducers happen to respond to nutriment and noxiousness serves to explain the adaptive functionality of the bacterium system, but does not constitute what is being represented by that system — the bacterium does not know anything about food or poison or first derivatives. It just swims or tumbles.

A frog's world is much more complicated, but the basic points remain the same. A frog can differentiate a tongue-flick-*at*-a-point followed by eating opportunity from a tongue-flick-*between*-two-points followed by eating opportunity from a dive-into-the-water situation, and so on. The fact that the first differentiations tend to occur with respect to flies, the second with respect to worms, and the third with respect to birds of prey will, as with the bacterium, help explain how and why these particular functional relationships are adaptive for the frog, but they do not in themselves constitute the representational contents for the frog. The frog tongue-flicks and eats, or it dives; it does not represent flies or worms or birds of prey from which it infers the proper behavior of tongue flicking or diving. What the frog *represents* are various tongue-flicking-and-eating situations, among others. Error is constituted if, for example, the internal states corresponding to eating do not follow.

The human world, of course, is enormously more complicated. The interactivist contention is that these same principles still hold, nevertheless. It is not at all obvious how the interactive approach could account for many phenomena of human epistemology and phenomenology. Perception, language, rational cognition, imagery, and consciousness are among the apparently problematic phenomena to be addressed. We will briefly outline the interactive model for some of these

phenomena for use in later discussions, but filling out the interactivist programme must be left for elsewhere (for example, Bickhard, 1980a, 1980b, 1987, 1992a, 1992c, 1993a, in press-a, in preparation-a, in preparation-b; Bickhard & Campbell, 1992; Bickhard & Richie, 1983; Campbell & Bickhard, 1986, 1992a).

This is a much abbreviated presentation of the central representational model of interactivism, and it does not begin to address the consequences of the view for standard modeling approaches nor any of its own programmatic developments. What can be noted from even this brief introduction is that functions of implicit definition, differentiation, and selections among further processing *cannot* in themselves constitute encodings. And they can certainly not be foundational encodings for two reasons: 1) because the representational content is *subsequent* to the elements, not constitutive of them, and 2) because the representational content is intrinsically distributed in the *organization* of potential processing selections, and is not necessarily localized or atomized in any *element* whatsoever.

Logically, then, it must either be denied that these functions have any relevance to representation at all, or it must be conceded at least that encodings cannot constitute the *essence* of representation, for here are representational phenomena that cannot be rendered in terms of encodings at all. Once this concession is made, it then becomes a programmatic issue whether or not interactivism can subsume such encodings as do exist.

This outline presents only the most basic core of the interactivist explication of representation (Bickhard, 1993a). The general interactive model, however, is a programmatic approach to all epistemic and mental phenomena, and has in fact been developed in a number of directions. Because a few additional aspects of the general interactive model will be needed for later points of comment and comparison, they will be outlined here. In particular, we will take a brief look at the evolutionary foundations of the model, the general approach to perception, a constructivist consequence for development, and the basic nature of the model of language.

EVOLUTIONARY FOUNDATIONS

The evolutionary foundation of interactivism consists of a sequence of knowing, learning, emotions, and reflexive consciousness that form a trajectory of macro-evolution. Knowledge is explicated as

being constituted in the capability for successful interactions (Bickhard, 1980a; Krall, 1992), as being intrinsic in any living system, and as inherently constituting interactive representations. Each of the later steps in the sequence — learning, emotions, and consciousness — is explicated in terms of specific changes in the system organization of the preceding step, and each is shown to constitute an increase in the adaptability of the resultant system. In that sense — because each arises from a change in the preceding, and each increases adaptability — the knowing, learning, emotions, and consciousness hierarchy is shown to be a potential macro-evolutionary sequence. Human beings are heirs of this evolutionary sequence, and knowing, learning, emotions, and consciousness form part of their innate potentiality (Bickhard, 1980a, in preparation-b; Campbell & Bickhard, 1986).

SOME COGNITIVE PHENOMENA

Perception

Perception is commonly construed as the first, essential, step toward cognition and language. *Knowing* interactions are foundational for interactivism, not perception. Simple living systems — such as paramecia — are successful, though primitive, knowers without any differentiated perception at all. Perception in the interactive view is, in a broad sense, simply the modulation of ongoing interactive activity by specialized subforms of interaction. In a narrow sense, perception *is* those specialized forms of interaction — specialized for their function of *detection* in the environment, rather than for functions of transformation and change. In higher organisms, of course, certain modalities of such detection-specialized forms of interaction have evolved anatomical, physiological, and neural specializations as well. The basic ontological character of perception as a specialized form of interactive knowing, however, is not altered by such substrate specializations (Bickhard & Richie, 1983).

In order to provide a sense of how differently such phenomena as perception can appear from within the interactive model, we will elaborate here on the interactive model of perceptual phenomena. Specifically, we will look at the interrelationships among notions of the *situation image, apperception,* and *perception.*

To begin, note that it will be functionally advantageous for a complex interactive system to construct and maintain an organization of action indicators. These would be indicators of all further interactions of

the system that are potentially available in the current situation, and of all still further interactions that might *become* potential if particular interactions were engaged in first. The organization of such indications of interactive potentiality, and of potentialities contingent on yet other potentialities becoming actual, is called the *situation image* (Bickhard, 1980b). An indicator, say **I**, in a situation image — possibly itself set directly by some differentiating interaction outcome, perhaps set by some more complicated process — indicates the possibility of particular further interaction types, particular further procedures, say $\mathbf{P_1}$ and $\mathbf{P_2}$. If $\mathbf{P_1}$ is engaged in, it will yield one of its possible outcomes, say **J**, **K**, or **L**. Therefore, the initial indicator **I** indicates the possibilities of, among other things, (creating via $\mathbf{P_1}$) one of the indicators **J**, **K**, and **L**. A situation image, in general, is constituted as vast webs of such functionally indicative relationships (Bickhard, 1980b). The situation image is the system's knowledge of what interactions can be performed, both proximately and with appropriate preparation, and, therefore, among which it can select in the service of goals.

The term "situation image" carries unfortunate connotations of encodings that need resisting. It is difficult to find terms concerning representation that are not already implicitly associated with encodingism, simply because encodingism dominates all presuppositions in this area. The interactive situation image, however, is no more than an organization of functional indicators, of the sort constructed by differentiator final states, and that might be constructed on the basis of those final states — for example, constructing a single (organization of) indicator(s) on the basis of many mutually context dependent indications.

A situation image is a primary resource for system interaction, and, as such, it repays considerable effort devoted to constructing, filling out, maintaining, and updating it. The process of such maintenance, updating, and elaborating or "filling out" is called *apperception.* It consists of the ongoing processes of the construction and alteration of the indicators constituting the situation image on the basis of the already constructed situation image and on the basis of new interaction outcomes. The elaboration process explicitly constructs indications of interactive potentiality that are implicit (perhaps in a complex and context dependent manner) in the already existing situation image and new interaction outcomes. Such elaboration will occur with respect to spatially extended implications — e.g., the unseen backs and sides of objects, as well as unseen rooms next door, and so on — temporally extended implications

— e.g., the expectable proximate and non-proximate future consequences of actions of the agent or of processes in the environment — and of various sorts of conditionals — e.g., if such-and-such an interaction is performed, then these other interactions become available as proximate potentialities. Various sorts of organizations within the situation image constitute our familiar representational world of objects in space and time, causally interconnected, and so on (Bickhard, 1980b, 1992c).

All interactions of the system will change some things about the world and will depend on certain conditions in the world in order to function successfully. This is simply a consequence of the physicality of interactions. Interactions that depend on certain conditions and that do not change those conditions in the course of the interaction can be used as detectors of those conditions (though not thereby as representers of those conditions). The detection of such implicitly defined conditions is the basic function that has been outlined for interactive differentiators.

The apperceptive updating of the situation image is based on both the ongoing situation image and on ongoing interaction outcomes. In the latter case, it is based on both detection functions and on transformational functions of those interactions, depending on which is most salient or learned, and perhaps on both. Some sorts of interactions are engaged-in almost exclusively for the sake of their detection functions, for the sake of their indications concerning future interactive potentialities, rather than for their own potentialities for changing the situation. In a broad *functional* sense, such sorts of interactions constitute *perceptual* interactions.

Some sorts of perceptual interactions, in turn, have shown themselves to be sufficiently important that evolution has developed physiologically and neurally specialized subsystems that are dedicated to these interaction types. These specializations have been with respect to various modalities of perception that provide to that species important information for the apperceptive updating of the situation image. Physiologically specialized, modality specific subsystems for apperceptive interactions, such as for vision, hearing, and so on, constitute the paradigms of perception. The broader functional sense of perception, however, will include such phenomena as apperceiving the environment via the tapping of a blind man's cane, sonar, radar, and so on. When detection interactions transcend any such physiological specialization, such as, for example, the brown ring test for iron in qualitative chemical analysis, we tend to not call them "perception," even

though they serve precisely the same function, minus evolutionary specializations. An evolved chemical test for iron that was specialized in the nervous system *would* be called perceptual.

Perception in this view is "just" a special sort of interaction engaged in for the purpose of apperceptive maintenance of the situation image. It is no more the only input to cognition than are the outcomes of *transformational* interactions. And perceptual interactions do not *yield* the situation image; they ground ongoing apperceptive *modifications* of it. Perceiving is ***not*** the processing of inputs into perceptions (Bickhard & Richie, 1983). In that sense, perception is not a matter of input at all, but, rather, an interactive modulation of situation image knowledge concerning further potential interactions. This view is quite different from standard encoding models of perception.

Learning

Interactivism also imposes distinct logical constraints on models of *learning*. Encodings are epistemic correspondences with the world. Consequently, it has classically been tempting to construe the origin of encodings as resulting from some sort of impression of that world on a receptive mind. The classic waxed slate, or *tabula rasa*, is the paradigmatic form. In contemporary work, this takes the more sophisticated form of *trans*duction for momentary patterns or elements in the world, and *in*duction for temporally extended patterns in the world.

The interactive representational relationship, however, is not a structural or correspondence relationship at all. It is a functional relationship of interactive competence for the potential interactions available. Both because it is a functional relationship, and not a structural relationship, and because it is a relationship with interactive *potentialities* of the world, and not with actually present actualities in the world, interactive representations are not logically capable of being passively impressed from the world on a receptive mind. Interactive representations must be *constructed* within the epistemic system, and then tried out for their interactive functional properties in a variation and selection process. The specifics of that process are, of course, deep and complex, but the basic point is that interactivism logically forces a constructivism of learning and development (Campbell & Bickhard, 1986; Bickhard & Campbell, 1988; Bickhard, 1988a, 1991c, 1992a, in preparation-a).

Furthermore, there can be no assurance that such construction of system organization will be correct. Any such assurances can only be based on prior knowledge of what will work, of what is correct. The origins of such knowledge is precisely what is at issue. Fundamentally, new knowledge must be constructed via some sort of variation and selection constructivism, a constructivism that is in the limiting case non-prescient, and that can account for the construction and use of prior heuristic knowledge in those circumstances when such prior knowledge *is* available. Interactivism forces an evolutionary epistemology (D. Campbell, 1974; von Glasersfeld, 1979, 1981).

There are many kinds of constructive processes that would constitute evolutionary epistemological systems of varying power. Bickhard (1992b) differentiates between, for example, the following:

- *simple* constructive processes that are always dealing with the same constructive materials in every new learning situation;
- *recursive* constructive systems, that can make use of previously constructed system organizations as "units" in new constructive attempts; and
- *meta-recursive* constructive systems that can, in addition, recursively construct in a variation and selection manner new constructive procedures, new procedures for learning and developmental constructions.

The move to topological dynamics (see below, and Bickhard & Campbell, in preparation) introduces still further complications. Such differentials of constructive power *within* an evolutionary epistemology will not generally be relevant to the issues discussed in this book.

The inadmissibility of prescience applies not only to the evolutionary and developmental constructions of system organization, but also to the *micro*genetic constructions of particular interactions and of apperceptive processing. The basic point is that, in order for a system to know precisely which interactions will function in what way — a visual scan interaction, for example — the system must already know what the environment is. Yet perceptual interactions are precisely what is dedicated to the differentiation of what that environment is. At any moment, we do in fact have vast prior knowledge of our immediate environment — knowledge based on prior interactions with this environment and prior encounters with the world in general. Consequently, the trial and error, variation and selection character of even

perception and apperception is not so clear to us. Most of the time we do have good foreknowledge (Bickhard, 1992a). In waking up in strange circumstances, however, or in difficult-to-perceive situations, the trial and error character of even such micro-genetic processes becomes felt. We try various perceptual interactions and various apperceptive interpretations to find out which will work. This variation and selection character of apperception also shows up importantly in the apperceptive understanding of linguistic utterances (Bickhard, 1980b; see below).

Language

Language is standardly construed as some form of encoding of mental contents, which in turn are construed as encodings derived from the basic encodings of perception. Interactivism undermines that sequence at every step. Simply, there are no basic mental or perceptual encodings for language to recode. Language thus takes on a quite different — an interactive — character.

Briefly, language is a special form of interaction, differentiated by the object with which it interacts. The basic intuition is that language is a conventionalized means for the creation, maintenance, and transformation of social realities. Social situations, then, are the special object of language interactions, and utterances are operations upon them. Some of the special properties of language derive from the special properties of social realities (Bickhard, 1980b), but several of the more striking differences from standard approaches already emerge with just the operative character of language, before the special social *object* of operations is taken into account.

For example, an utterance operates on an initial social situation and transforms it into a resultant social situation. The result of the utterance will, in general, be as dependent on the initial context of the utterance as on the utterance itself. Language, in other words, is in this view *intrinsically context dependent*. Further, utterances operate on social situations, which intrinsically involve representations of that situation and its participants, but the utterances themselves are *operations* on such representations — they transform initial ones into resultant ones — and are not representational themselves.

Language, then, is fundamentally not encodings. In fact, language is fundamentally not representational at all. Just as an operator on numbers is not itself a number, so an operator on (social organizations of) representations is not itself a representation. Such consequences for

language are partially acknowledged in recent distinctions between content and character, with *character* corresponding to utterance operational power (Fodor, 1987, 1990; Kaplan, 1979a, 1979b, 1989; Richard, 1983).

Among other consequences, this point scrambles the standard distinctions between syntax, semantics, and pragmatics. Syntax is typically taken to be the study of well formed encodings; semantics is the study of the encoding relationships; and pragmatics is the study of how such encodings are used. This syntax-semantics-pragmatics framework for the study of language is presumed to be theory- and programme-independent, but in fact it is committed to encodingism. In the interactive view, the phenomena of language do not fit together in the way that this framework presupposes (Bickhard, 1980b, 1987, in press-a; Bickhard & Campbell, 1992; Campbell & Bickhard, 1992a). The intrinsic meaning of an utterance, for example, is operational, functional, pragmatic — and *not* representational — while utterance can be *used* to create various representations with truth values.

There are some interesting constraints between the interactive model of representation and the model of language to which it has given rise. Corresponding to the distinction between encoding and interactive models of representation, there is a distinction between *transmission* and *transformation* models of language (Bickhard, 1980b). *Transmission* models of language construe utterances in the classical mold as (re-) encodings of mental contents that are transmitted to other minds, where they are decoded into mental encodings that constitute understanding. *Transformation* models construe utterances as operators, as transformations, on social realities. (It should be clear that transformation models of language have little to do with transformational grammars. In fact, there are deep incompatibilities, not the least of which derives from the basic encoding presuppositions of transformational grammars.)

Transmission models of language are straightforward extensions of encoding models of representation: utterances are just another step of encodings. Transformation models, similarly, are extensions of the interactive model of representation. Here, utterances are a special kind of interaction that transforms the world, transforms social realities in this case. In addition to these natural affinities, there is at least one strict incompatibility: transmission models of language cannot be built on interactive models of representation. The basic reason for this is that interactive models of representation do not provide the necessary

elements within individuals, and commonalities of organizations of such elements between individuals, that are required for utterances to begin to be construed as encodings of mental contents. There are further moves in the argument to which this claim leads, which we will not recapitulate here, but the basic point of a severe incompatibility should be clear (Bickhard, 1980b, 1987).

There is also an incompatibility in the other crossed direction, between transformation models of language and encoding models of representation, but it is not as logically strict. It is at least superficially conceivable that representation could consist of encoding elements, while utterances consist of actions that operate on, that transform, those encoding elements. The fit, however, is awkward and forced, and it leaves many difficult, perhaps impossible, problems. For example: How are utterances *as transformations* produced from encodings? What is the object of utterances as transformations? How are utterance-transformations understood by an audience so as to alter the audience's mental encodings? If utterances are themselves construed as encodings, these questions seem to have, at least in principle, clear sorts of answers. Plausible approaches to the questions in the case of the forced hybrid are far from clear.

The general point, then, is that interactivism as a model of representation undermines standard conceptions of language — transmission models of language cannot be combined with interactive models of representation — and it logically forces something like the transformational model. Given interactivism, utterances must be some sort of interaction; the question is: What sort? The converse constraint is not quite as strong, but it is still powerful: transformation models of language do not immediately logically force an interactive model of representation, but their incompatibility with encodingism is, nevertheless, strong.

Interactivism's focus on language as a functional activity is partially convergent with Wittgenstein's notion of meaning as use. Wittgenstein's reliance on "criteria" to connect a purported representational function of language to the world, however, commits him, in spite of his own criticisms of his earlier *Tractatus* encoding model, to an encoding conception of representation (Bickhard, 1987). Interactivism's intrinsic context dependence, considered from the perspective of a historical text, yields the hermeneutic context dependency — i.e., the interpreter's historically located initial

understanding constitutes the context that will be transformed by the text. Similarly, the impossibility of utterances being encodings, even encodings of operations, necessitates that the interpretation and understanding of them is intrinsically an open variation and selection problem solving process. Interpretation is a variety of apperception. This process will be habitualized and automatized in varying degrees depending on the familiarity of the situation and operations (text) involved. The iterations of attempts and approximations involved in solving the open problem of interpretation yields the hermeneutic circle (Gadamer, 1975; Heidegger, 1962; Ricoeur, 1977). As mentioned above, however, hermeneutics in its standard form is committed to a linguistic idealism, which interactivism challenges as being itself a version of incoherent encodingism.

Most of the interactive model is missing from this account, but, along with various further elaborations in later discussions, this should suffice for the analyses at hand. Most fundamentally, interactivism is part of an attempt to replace standard *substance and structure* ontological approaches to mental phenomena with strict *process* ontologies. Psychology, philosophy, linguistics, and Artificial Intelligence alike are replete with such substance and structure ontologies; they are still embedded in the ontological equivalents of phlogiston, magnetic fluid, vital fluid, and other such substance approaches. Most sciences have long ago understood the fundamental inadequacy of such substance approaches, and have abandoned them for more adequate process ontologies; sciences of the mind and of mental phenomena, however, have not. These discussions have focused, and will continue to focus, primarily on the implications of a shift to process ontologies for representation and language, but the general ontological psychology approach, of which interactivism is a part, attempts to go far beyond those (Bickhard, 1991c, 1993a, in preparation-a, in preparation-b; Bickhard & Christopher, in press).

8

Implications for Foundational Mathematics

TARSKI

One way to understand the scope and depth of the encoding critique and the interactivist alternative is to consider the two foundational forms of mathematics for all of Cognitive Science: Tarskian model theory and Turing machine theory. The encodingism critique and the interactive alternative invalidates both approaches. Tarskian model theory is the historical ground and general form of almost all contemporary approaches to semantics — whether linguistic or cognitive — and such "semantics," of course, is the general approach to issues of meaning and representation (Eco, Santambrogio, Violi, 1988; Field, 1980, 1981; Barwise & Etchemendy, 1989; Nilsson, 1991).

Encodings for Variables and Quantifiers

The critical point here is simply that Tarskian model theoretic semantics is no more than a sophisticated and formalized encoding model. The brilliant contributions that Tarski made to the basic encoding intuition included showing how that intuition could be formalized for variables and quantifiers — not just for objects, properties, relations, and logical connectives — and showing how to rescue the encoding notion of "truth" from intrinsic paradox. Tarskian model theory, however, only renders the "semantics" of one language in terms of the unanalyzed, but used, semantics of another language — the language in which the *model* is stated. It addresses only the semantics of derivative encodings. It does not, and can not, provide a semantics for any foundational, logically independent, language or representational system. The additional power introduced by moving to model theoretic *possible worlds* semantics, of

course, does not alter this basic point at all (Bickhard & Campbell, 1992; Campbell & Bickhard, 1992a).

Model theory, then, does not provide a way for encodingism to solve the problem of emergent representation — it does not provide a way in which representation can emerge out of phenomena that are not themselves already representational. It cannot, therefore, offer a complete theory of representation or meaning. On the other hand, an interactivist approach — with its principles of differentiation and selection, operation and transformation — can capture the power of model theory and logic. There are at least two ways to approach this point. The first way is to note that the correspondences between language and model that are formalized as mappings in model theory can instead be understood as differentiations and selections (Resnick, 1981). The second is to recognize the formal equivalency of approaches to logic based on operations instead of mappings. (Algebraic logic explores and develops such an approach. Interestingly enough, this is an approach to which Tarski has also made fundamental contributions; Bickhard & Campbell, 1992; Campbell & Bickhard, 1992a; Craig, 1974; Grandy, 1979; Henkin, Monk, & Tarski, 1971; Quine, 1966b.) In either sense, interactivism can capture the power of standard approaches, but not vice versa. In particular, standard approaches cannot model representational emergence, and they cannot solve or avoid the incoherence problem.

Because Tarskian model theory, or some variant or derivative of it, forms the ground for virtually all contemporary approaches to representation — linguistic or cognitive — the inadequacy of model theory to solve or avoid the basic problems of encodingism constitutes a serious deficiency. The fact that model theory just *is* a sophisticated encodingism implies that Artificial Intelligence and Cognitive Science themselves have no alternative to encodingism, no solutions to the incoherence problem or the problem of emergence.

Tarski's Theorems and the Encodingism Incoherence

Tarski's theorems, however, not only do not *solve* the problems of encoding incoherence. In fact, they *exemplify and demonstrate* those problems. They provide additional, and mutually illuminating, perspectives on the fundamental encodingism incoherencies.

The organization of Tarski's proofs concerning Truth predicates has the following form (Field, 1980; Martin, 1984; Tarski, 1956). He proved that in any language **L'** that is "adequate to the semantics" of a

primary language **L**, the Truth predicate for **L** could be constructed. If the languages are identical, **L = L'**, then that Truth predicate could be used to construct the semantic paradox of the liar within **L**. If **L'** is a *meta*-language for **L**, then that construction of the liar paradox is blocked. Any language, then, that is "adequate to its *own* semantics" is thereby logically inconsistent by virtue of the combinatoric constructability of the liar paradox. "Adequate to its own semantics" in this context basically means "able to capture its own encoding correspondences — its own semantics in an encoding sense." That is, "adequate to its own semantics" means "able to encode its own model theoretic semantics." In these terms, any language that can supposedly capture its own encoding representational relationships to its semantic domain, to the "world" outside of itself, is intrinsically inconsistent. A meta-language, however, is capable of capturing these encoding relationships without inconsistency.

Both aspects of Tarski's theorems are relevant to the encodingism critique. The inconsistency of a language that supposedly captures its own semantics is itself a manifestation of the encodingism incoherence problem, and the manner in which such a semantics *can* be consistently represented from a meta-language is a formalization of an observer semantics. Jointly, then, the two sets of theorems demonstrate the necessity of an observer semantics in order to make good on an encodingism.

Representational Systems Adequate to Their Own Semantics

First, consider the impossibility for an encoding language to capture its own semantics. Tarski's theorems show that the definition of Truth for **L** requires that the encoding semantics for **L** be captured. To assume that the encoding semantics for **L** can be captured *from within **L*** yields inconsistency. From the interactive perspective, to assume that an encoding semantics could be captured from within that encoding system is false — to make such an attempt encounters the encoding incoherence and fails. Thus, if **L** *does* presume to capture its own semantics, that constitutes a false assumption within **L**, and the consequence of logical inconsistency of the encoding system **L** follows necessarily and expectably.

That is, to suppose that an encoding system **L** is adequate to its own semantics — can represent its own encoding representational relationships — is *precisely* to suppose that **L** can bridge the incoherence

of encodingism and provide representational content to its foundational encodings. It is to assume that the language **L** can escape the solipsism of encodingism — that it can provide an observer semantics from outside of itself onto its own epistemic relationships to its represented world. The incoherence argument shows that this presupposition is intrinsically false, and, therefore, that any logical system making that presupposition — to be able to represent what it in fact cannot represent — will be thereby incoherent, and subject to inconsistencies.

Note, in this regard, that to assume that **L** captures its own semantics yields that **L** is inconsistent, which, in turn, destroys any meaningful sense in which **L** could be said to capture its own semantics in the first place — any sentences in **L** supposedly encoding the semantics of the encodings of **L** could just as validly be replaced by their negations. In other words, there will *be* no coherent semantics in an encoding language **L** that presumes to capture its own semantics.

Observer Semantics

On the other hand, the semantics of **L**, and the Truth predicate for **L**, *can* be consistently defined from within a meta-language for **L**, **L'**. The semantics of the meta-language **L'** suffices to define the Truth predicate of **L** if and only if it is adequate to the semantics of **L**. But this involves the semantics of **L'** being able to represent, to encode, the semantics of **L**; this, in turn, involves **L'** being able to encode both the elements of **L** *and* the "elements" of the "world" that are encoded by those elements of **L**. In other words, **L'** must be able to define and construct the encodings that constitute the semantics of **L**, and, to do that, **L'** must have independent semantic, representational, perspectives on both the language **L** *and* on the "world" of **L** — the semantic domain of **L**. But this implies that, in order for the meta-language **L'** to be adequate to the semantics of **L**, and thereby able to construct the Truth predicate for **L**, that meta-language must capture an *observer semantics* for **L**. In other words, Truth cannot be consistently defined within this framework *except* through the use of the *already existing* representational power, the semantics, of an observer meta-language.

Consistent with the interactive critique of encodingism, then, the encoding semantics for **L** can be captured, and can *only* be captured, from *outside* the encoding system **L** itself. It requires independent perspectives on both the encoding language **L** *and* on its domain of representation, and such a perspective *cannot* exist from within **L** itself — it requires an

external observer perspective. The encoding semantics of **L** is the *only* perspective that **L** has on its own semantic domain, and to attempt to define that semantics *in terms of that same semantics* is *precisely* the incoherence circularity. Such independent external perspectives on both **L** *and* on its semantic domain, however, can be formalized in a *meta*-language for **L**, **L'**.

Tarski's theorems, then, reflect further aspects of the fact that encodingism cannot constitute an adequate approach to representation. Any attempt to do so yields inconsistency, and avoiding the inconsistency requires reliance on the semantics of a meta-language, a formal stand-in for an observer. As noted earlier, it is precisely such ultimate observers, such ultimate semantics, that we would like to understand. Formulating the encoding semantics of one language **L** in terms of the semantics of a meta-language **L'** is useful and powerful for many purposes, but it does *not* constitute a model of semantics or representation per se. It only shifts the unknown from **L** to **L'**. We already know that encodings can be defined in terms of other encodings, but it is the nature of representation per se that is ultimately at issue. Tarski's theorems provide one more aspect of the impossibility of understanding that from within an encodingism.

Tarski's theorems are important both for what they show can be done and how to do it, and for what they show cannot be done. They show that in order to get a consistent encoding semantics for a language, a meta-language semantics must be used. They also show that to presume that an encoding language can capture its own semantics is intrinsically inconsistent.

Truth as a Counterexample to Encodingism

Tarski's theorems about Truth provide a counter-example to encodingism: Truth cannot be consistently defined within the syntactic combinatorics of encodingism. Conversely, the interactivist incoherence argument provides a different perspective on the import of Tarski's theorems: the inconsistency of a language presumed to be adequate to its own semantics is an aspect of the presupposition of an incoherency, the foundational encodingism incoherency. Still further, an encoding semantics *can* be coherently captured from an appropriate external observer perspective, but it then provides no explication of encoding representation per se — it "simply" uses one unexplicated encoding system to represent characteristics of some other encoding system.

The inconsistency of presuming that an encoding language is adequate to its own semantics is readily interpretable from an interactive perspective — it is simply a formal manifestation of the incoherence of strict encodingism. In this respect, the interactive critique provides an explication of the difficulties regarding Truth and semantics that are demonstrated by Tarski's theorems. This point may generalize: the semantic paradoxes in general, not just the liar paradox, (and, arguably, the set theoretic paradoxes as well — though we will not develop the arguments here) involve similar presuppositions that a representational system can make good on its representational encoding correspondences — can cross the gulf of the incoherence of foundational encodings, and escape the resultant solipsism. Although such paradoxes in general involve self-referentiality, it is clear that self-referentiality *per se* does not yield paradox. We suggest that the problem, at least in many cases, derives most deeply from that particularly circular form of self-referentiality that assumes that an encoding system can claim " '**X**' represents whatever it is that '**X**' represents" and get away with it (cf. Barwise & Etchemendy, 1987; Gupta & Belnap, 1993; Herzberger, 1970; Martin, 1984; Priest, 1987; Sheard, 1994; Visser, 1989; Yaqub, 1993).

A general moral of this story of Tarski's theorems concerning Truth is that, not only is the assumption of the adequacy of encodingism a false assumption — it cannot capture, for example, Truth — it is an assumption that can yield further deep logical errors, errors that are not easy to discover or understand. That is, encodingism is not only wrong, it is also conceptually dangerous.

TURING

As Tarskian model theory provides the ground for contemporary approaches to semantics, so does Turing machine theory provide the grounds for contemporary approaches to process. Just as few models will be stated directly in terms of Tarski's model theory, so are virtually no models constructed directly in terms of Turing machine theory — more tailored languages and models are used. But being more tailored, for all the importance that can and at times does have, does *not* imply being fundamentally more powerful. Turing's thesis, in fact, states that a Turing machine is capable of any formal process that can be performed. This thesis is not capable of proof, but it is capable of disproof, and more than half a century of mathematical research has yielded the consensus conclusion that the thesis is true (Rogers, 1967; Cutland, 1980). Turing

machine theory, then, is a natural locus for in-principle discussions of computation (Hopcroft & Ullman, 1979; Minsky, 1967).

Semantics for the Turing Machine Tape

In spite of the success of Turing's thesis, there are two fundamental problems with Turing machine theory as it is implicitly involved in Artificial Intelligence and Cognitive Science. These problems arise because of a usually implicit, though sometimes explicit, uncritical extension of Turing's thesis that is universal in these areas: Turing's thesis is strictly stated in terms of *formal* processes, but it is treated as if it were true for *all* processes, or at least for all cognitive processes (Levesque, 1988; Turing himself may have entertained such extensions: Hodges, 1988). The two problems arise directly from the two senses in which this is an extension of Turing's actual thesis. The first is simply that Turing was attempting to capture operations on *uninterpreted* symbols — symbols without meaning or semantics, with no representational power. To extend it to cognitive activity, then, and thereby assume its adequacy to phenomena of representation, is to populate the Turing machine tape with *interpreted* symbols, not uninterpreted symbols. The promissory note of interpretation, of course, is supposedly filled with Tarskian model theory. This extension of Turing's thesis, then, founders on the direct incorporation of model theoretic encodingism.

Sequence, But Not Timing

The second problem with the extended Turing's thesis is that *formal* process is process in which only *sequencing* of operations, not their timing, is of relevance. The focus for Turing was the logic of mathematical proof, and he was concerned with what the steps in constructing a proof might be and the sequences in which they might occur (Herken, 1988; Hodges, 1983). It is, in fact, such formal steps and their bare sequence that *constitute* a formal process in the sense that Turing meant it, and to which Turing's thesis refers. The actual *timing* of these steps was irrelevant to the concerns that Turing was focused on, and is not and cannot be formalized within Turing machine theory, or any of its equivalents.

Turing machine theory, then, cannot accommodate timing considerations, and, therefore, cannot model temporally critical processes. It is powerless with respect to temporal coordination, for example, or

temporally critical aspects of an interaction with an environment. The most extended version of Turing's thesis, then, that assumes it for *all* process, is simply false. Turing machines can handle only sequence, not timing (Bickhard & Richie, 1983, p. 90; van Gelder & Port, in press).

Clocks. An apparent rejoinder to this would claim that all that is needed to handle timing issues is the introduction of a clock into the model. A clock, or some functional equivalent, in fact, is exactly what is needed. *But there is no formal way to model a clock in Turing machine theory*. The steps of a formal sequence could be wildly unequal in terms of actual timing — one second for the first step, a century for the second, fifteen nanoseconds for the third, etc. — and the logic of Turing machine theory would not be affected at all. There is nothing akin to an oscillator in Turing machine theory, and no possibility of constructing one, and, therefore, no possibility of a clock.

It is certainly the case that any actual construction of a physical instantiation of a Turing machine will necessarily be concerned with the timing of the actual physical instantiations of the formal relationships involved in the theory, and a clock is a sensible and handy way to solve those instantiation timing problems. But such clocks — as in contemporary computers — are *engineering* introductions, not formal or theoretical introductions. Computers are Turing machines engineered with clocks to drive the sequences of steps of processing. Clocks are designed-in at the engineering level in order to in fact instantiate the formal machine, but there are still no clocks, no oscillators, in the theory itself. However much, then, that clocks can and must be designed into an *instantiation* of a Turing machine, this does not affect the fact that Turing machine *theory* cannot model clocks or oscillators, and, therefore, cannot handle issues that involve timing. Similarly, neither can any languages that are formally equivalent to (or weaker than) Turing machine theory.

Computers are a practical advance over abstract Turing machines in that they do contain clocks, and their lowest level processing steps do (or can) manifest equal temporal intervals. They allow, then, programming for real time activities by taking into account the clock time. But, to reiterate, this is an engineering level introduction, not a theoretical account, and does not contribute to the theoretical understanding of necessarily real time interaction. Further, it is the *programmer* who takes into account the clock time, not the computer, and we find, again, a user semantics, not a for-the-machine semantics. This point introduces a second version of the rejoinder.

This variant of the rejoinder would be to introduce a clock not in the instantiation *of* the formal system, but as a generator of formal inputs *to* the formal system — clock ticks as inputs. These inputs, however, will either be formally uninterpreted, or they will be taken to be interpreted symbols, representing time units. If they are formally uninterpreted, they do not add to the theory — they will simply be a sequence of identical empty symbols which *happen* to be input at equal time intervals, a fact totally beyond the competence of the theory to model or take into account. If they are interpreted as symbols, we must ask how the system knows what they represent, how they are interpreted, and then all of the incoherencies of encodingism are encountered again.

Such inputs, of course, might be extremely useful to a *programmer* of such a system, but such usages involve a user semantics, not a system semantics. Such usages do not involve any extension of Turing machine theory at all. Rather, it is the programmer who must keep track of the "equal time interval of X-many milliseconds" significance of those inputs, without any such significance being captured in the theory itself, and, furthermore, without any such significance being captur*able* in the theory itself. As in the instantiation case, timing considerations can, and for some purposes must, be introduced *in addition* to the Turing machine theory (or programming language equivalent), but they cannot be captured *within* the theory or language itself. There is no way even for the mathematics to represent that the clock inputs are of equal time intervals, and certainly not to represent, in a *system* semantics, that they are of some particular length. As before, timing is fundamentally irrelevant to Turing machine mathematics, and, correspondingly, Turing machine mathematics is fundamentally incompetent with respect to timing.

Is Timing Relevant to Cognition?

A second rejoinder might be to acknowledge that the fully extended version of Turing's thesis to *all* process *is* invalid, but to still uphold it for *cognitive* processes because cognition *is* just formal process, and timing issues are not relevant. Equivalent claims might be that cognition is just operating on pointers, or on formal data structures, and so on. This is, in fact, the form in which an extension of Turing's thesis is usually presupposed or proposed. The interactive model, however, perforce implies that this weaker extension also is invalid.

In particular, if the interactive model is valid, then all representational phenomena, even, ultimately, formal phenomena, are intrinsically grounded in actual interactions between actual systems and actual environments in real time, with timing considerations generally playing a critical role. From visual scans to walking, interaction requires timing — inherently. The interactive notion of representation is fundamentally dependent upon that of interactive success, and the goals and criteria with respect to which "success" can be defined. Interactive success, in turn, is fundamentally dependent on getting the timing of the interactions, not just the sequencing, right. Issues of timing, then, are foundational to issues of representation, not secondary adjuncts. Timing — *oscillators* — must be an integral part of the theory, not an engineering introduction underneath the theory.

Turing machine theory, as usually presupposed, then, not only *directly* incorporates encodingist Tarskian model theory in order to interpret the symbols on the machine tape, it is also *indirectly* committed to encodingism in that formal representation must be secondary, must be derivative encodings, since formal systems cannot capture the timing aspects that are essential to emergent, grounding, forms of representation. Conversely, a formal approach consistent with interactivism must involve oscillators and their interactions, or some functional equivalent, in the grounding ontology of the theory.

Transcending Turing Machines

Just as interactivism is capable of the power of model theory, but not vice versa, so also is interactivism capable of the power of Turing machines, but not vice versa. A simple in-principle way to demonstrate this is to point out that the oscillatory aspect of a interactive system ontology is *already* formally competent to Turing machine theory. One form of limiting case of modulation of one oscillator by another is modulation to the extreme of damping the oscillations of the second oscillator out of existence, of switching the second oscillator off — or of evoking oscillatory activity in the second oscillator, of switching the second oscillator on. That is, one form of limiting case of intersystem *modulation* relationships is intersystem *switching* relationships, and that is already enough to construct a Turing machine.

Note that the switching relationship already abstracts away from most of the inherent temporal properties of modulation relationships. Note also that modulations of fields of oscillation is deeply characteristic

of brain functioning. If interactivism is correct, then that characteristic of brain functioning is no accident of instantiation, but is intrinsic in the interactive character of representation (see the discussion of interactive architectures below). It is intrinsic in the sense that representation is an aspect of action and interaction — not just a functional *adjunct* to interaction — and action and interaction require timing, which inherently involves the ontology of oscillators. Conversely, any modeling approach that is adequate to representation will necessarily involve oscillatory temporal ontologies, not just engineering clocks.

To briefly mention two additional implications here, we note 1) that oscillations and oscillatory modulations are intrinsically continuous, and the space of their dynamics has intrinsic topologies, unlike the discrete algebras of formal systems; and 2) that fields of oscillators are intrinsically parallel and concurrent in their functioning and in their modulations. Still further, oscillations and their modulations can be superimposed, so that the "messages" among oscillators are intrinsically concurrent, unlike the discrete formal messages among parallel formal systems. Message hangup is not a threat in interactive modulatory systems. We will not pursue these implications further here, but wish only to note that the requirements of a formalization of interactive processes force the use of languages fundamentally different from those to be found in contemporary Cognitive Science, and, much more powerful languages (see Section IV).

III

ENCODINGISM: ASSUMPTIONS AND CONSEQUENCES

9

Representation: Issues within Encodingism

The discussion to this point has been primarily at the conceptual or programmatic level. We have been concerned with the foundational issues of encodingism and with some of the partial attempts to address them. We turn now to analyses of some representative approaches or projects *within* Artificial Intelligence and Cognitive Science to try to provide some more specific senses of the involvement and consequences of encodingism. The overview will focus on several themes in the recent history of the field. First, we address issues of representation, followed by language, then learning, and finally a discussion of connectionism and PDP.

The encoding assumptions of Artificial Intelligence and Cognitive Science are sometimes explicitly stated, but more commonly they are implicit. Encodingism is so *pre*supposed, so taken for granted, that it is often not stated or acknowledged at all. Encodingism, after all, appears to be all there is for understanding representation, so it is quite understandable that it would appear to need no separate statement or acknowledgement. When the general nature of representation is explicitly addressed, representation is at times simply asserted to *be* encodings (e.g., Palmer, 1978).

Within the general framework of encodingism, however, there are an unbounded number of variations on the basic theme, all having to do with the semantic nature and specifics of the elemental encodings; the presumed generation of encodings; the syntax of acceptable combinations; the relationships with systems that operate on encodings; the implementation of the encodings and systems; and so on. The unbounded variety of options available for addressing these issues means that they *must* be addressed, and, in fact, they are generally the direct focus of investigation and theorizing — while the encodingism

framework itself remains taken for granted (e.g., Melton & Martin, 1972; Neisser, 1967; Glass, Holyoak, Santa, 1979; Fodor & Pylyshyn, 1981; Bobrow, 1975; Barr & Feigenbaum, 1981; Rich & Knight, 1991; Posner, 1989; Rumelhart & Norman, 1985; Barr, Cohen, Feigenbaum, 1981-1989).

Even though encodingist assumptions remain in the background, research explorations within such a framework can, nevertheless, still encounter the limitations and incoherences of encodingism. That is, the problems that are *focal* in such work are the details of designing and implementing the encoding-computational systems, but the underlying programmatic problems of encodingism can interfere nevertheless. We begin by examining several projects that are squarely within the encoding framework, and point out some of the ways in which the intrinsic encoding limitations are manifested. We have selected these projects because they are well-known and representative — or (inclusive) because they provide illustrations of points that seemed worth making.

EXPLICIT ENCODINGISM IN THEORY AND PRACTICE

Physical Symbol Systems

Overview. In Physical Symbol Systems (Newell, 1980a), Allen Newell has attempted to define precisely what he considers to be the central concerns of Artificial Intelligence and Cognitive Science. He advances the notion that general intelligence is a property of *physical symbol systems*, a somewhat precisely stated version of familiar AI symbolic processing systems. This hypothesis was proposed in Newell & Simon (1972, 1975) and endorsed in Newell & Simon (1987) and Vera & Simon (1993, 1994). Newell argues (and we agree) that the Physical Symbol System Hypothesis "sets the terms" on which Artificial Intelligence scientists search for a theory of mind (Newell, 1980a, p. 136). As such, it is a compelling subject for an interactivist critique — demonstrating how such an influential notion within Artificial Intelligence is committed to encodingism reveals the foundational flaws within the Artificial Intelligence programme. In addition, we will discuss Newell's Problem Space hypothesis and the SOAR "cognitive architecture" of Newell and colleagues, a project consciously carried out following the Physical Symbol System Hypothesis, to illustrate how the foundational flaws of Artificial Intelligence weaken specific projects.

As a model of the workings of computers, we have no major objections to the Physical Symbol System Hypothesis. But, in claiming

that general intelligence is a property of such systems, the hypothesis makes claims about cognition more broadly, including representation. It is here that we find fatal flaws — the flaws of encodingist assumptions about the nature of representation.

A central notion in the Physical Symbol System Hypothesis for issues concerning representation is that of *access.* Access is a strictly functional relationship between a machine and some entity. Internal to the machine, such an accessible entity could be a symbol, an expression, an operator, or a role in an operator. Access basically means that the machine can operate on whatever it has access to — for example, retrieve a symbol, change an expression, and so on. **Assign** is an operator that assigns a symbol to some such internal entity, thereby creating access to that entity. Access to such an assigned symbol yields access to the entity to which that symbol is assigned. Assignment, then, creates a kind of pointer relationship that constitutes functional access.

The next major notion for our purposes is that of *designation*.

> *Designation*: An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y. (Newell, 1980a, p. 156)

In other words, "the process behaves as if inputs, remote from those [that] it in fact has, effect it. ... having X (the symbol) is tantamount to having Y (the thing designated) for the purposes of process P" (1980a, p. 156).

Such a remote connection is created "by the mechanism of access, which is part of the primitive structure [of the machine] ... It provides remote connections of specific character, as spelled out in describing **assign**" (1980a, p. 156). To this point, we have a description of various sorts of functional relationships and possibilities internal to a machine.

We next find, however: "This general symbolic capability that extends out into the external world depends on the capability for acquiring expressions in the memory that record features of the external world. This in turn depends on the **input** and **behave** operators" (1980a, p. 157). "**Input** ... requires its output symbols to reflect an invariant relation to the state of the external environment (via states of the receptor mechanism)" (1980a, p. 167).

And, finally: "Representation is simply another term to refer to a structure that designates:

> X *represents* Y if X designates aspects of Y, i.e., if there exist symbol processes that can take X as input and behave as if they had access to some aspects of Y" (1980a, p. 176).

A representation, then, is a representation by virtue of the fact that it designates what it represents, and it designates something insofar as it provides access to it. Again, as a model of the internal workings of a machine, this is largely unobjectionable. When it is extended to epistemic relationships between the machine and its environment, however, it fails.

Critique. Considered from an interactivist perspective, one of the most perspicuous characteristics of the physical symbol system is its severe incompleteness. For comparison, recall that interactive representation consists of three aspects:

- **Epistemic Contact**. Interactions with an environment terminate in one of two or more possible internal final states, thus implicitly differentiating the environment with respect to those possible final states. This is the *epistemic contact* aspect of representation — the manner in which interactive representations make contact with particular environments.
- **Functional Aspect**. Internal states or indicators, generally constructed with respect to dependencies on such final states, influence further system processing. This is the *functional* aspect of representation and is the only role representations can play within a system.
- **Representational Content**. Through influencing goal-directed interaction, which either succeeds or fails in achieving its goals, *representational content* emerges in the organization and functioning of a system as falsifiable implicit interactive predications about the environment. Representational content has truth value that is fallibly determinable by the system itself, not just by an observer.

The Physical Symbol System Hypothesis, in contrast, focuses primarily on an "intelligent" system having processes that operate on and transform internal symbol structures — expressions. This is an abstraction of the model of a computer program operating on a data structure. The Physical Symbol System Hypothesis is not a full statement of even the functional aspect of representation (though it gestures in that direction), because the focus is on the transformations of internal records rather than on the influence of internal states on further processing, and because that notion of transformations does not in any essential way depend on action or interaction. The further processing is, in general, merely more manipulations of internal "records."

Thus, even though the focus of the Physical Symbol System Hypothesis is primarily on functional characteristics, it is nevertheless incomplete even with respect to the functional aspect of representation. The physical symbol system definition emphasizes processes that generate new internal "representations" out of already present "representations." That is, the Physical Symbol System Hypothesis defines process in a manner that *presupposes* issues of representation — processes operate on "symbols" — instead of providing an account of the emergence of representation out of process (Bickhard & Richie, 1983). A model of the emergence of functional processes must be independent of issues of representation, because function is logically prior to representation, with the emergence of representation then modeled within that framework of functional processes (Bickhard, 1993a).

The Physical Symbol System Hypothesis has it backwards: it assumes that representation can be defined prior to process, and then that processes can be characterized in terms of their effects on representations. It does not recognize that the functional influence of internal states on further processing is the limit of what internal states can do or be, and that a model of representation must be consistent with that fact. And it does not recognize the importance for representation — for genuine symbols — of *interactive* processes at all. Consequently, there is nothing in this model that provides either epistemic contact or representational content. The core hypothesis, in fact, does not even address the issue of representational content.

The Physical Symbol System Hypothesis does, however, make a gesture toward epistemic contact in the notion that the operator **input** generates symbols that "reflect an invariant relation to the state of the external environment" (1980a, p. 167). Such an invariant relationship is taken to provide representation, designation, and access to that state of the external environment.

There is an ambiguity here between two different notions of "access." Internal to a machine, a symbol can provide access to some other entity by providing a pointer to it. Alternatively, one entity could provide a kind of access to another by virtue of the first entity constituting a copy or an isomorph of the second. In this case, the machine could function in ways sensitive to features of the "designated" entity simply because the first entity provided the same features as the second. An important property of designation — transitivity — fails to distinguish between these two possibilities: "if X designates Y and Y designates Z,

then X designates Z" (1980a, p. 157). Both pointer access and isomorphy are transitive.

External to a machine, however, the two possibilities are on quite different footings. Pointer access cannot exist to the environment in the sense in which it does internal to the machine: internal access is simply **assign**ed, and is a primitive in the architecture of the machine, e.g., in the hardware for memory retrieval. There is no such primitive for the environment. It might be claimed that pointers for the environment could be constructed that would permit retrieval via various actions and interactions with that environment, such as providing a spatial location of the designated entity. This is certainly correct, but the machine's interpretation of such pointers involves representational issues, and thus would be circular as a foundation for a model of representation. If the pointers are taken to be simply commands to the operator **behave** that accomplish the required actions for retrieval — without representational interpretation — then we have at best a control system that can arrive at various locations in accordance with internal controls. Issues of representation, including representation of whatever it is that is at the "designated" location, are not addressed by such a model.

Newell emphasizes the invariance of relationship between the internal "symbols" from the operator **input** and states of the environment. He does not present a pointer relationship for **input**. Such an invariance of relation to the environment is a general form of isomorphy or tracking or correspondence with that environment. This is also the kind of relationship emphasized, for example, by Vera & Simon (1993). These relationships too are quite possible and important. They provide possible control relationships between the environment and internal processes of the machine, such as a photocell opening a door, or a thermostat adjusting a furnace, or a pin-prick evoking a withdrawal in a flatworm, or a keystroke on a keyboard triggering various activities in a computer, and so on.

Such factual correspondences are crucial to effective and appropriate sensitivity of the machine or system to its environment. They provide the possibility of such sensitivity because they provide the possibility for control influences, control signals, reaching from the environment into the system — and, therefore, the possibility for the system to respond to those signals, to be controlled by those signals, thus manifesting the required sensitivity.

Such control signals, however, do not provide any representational relationships. They are factual relationships of correspondence or tracking that provide the possibility for control relationships of process evocation or other process influence. They might provide a minimal form of epistemic contact (they are a minimal — passive — version of an "interactive" differentiator), but they provide nothing toward representational content.

In particular, to assume that these internal states correspond to objects or entities in the world, and thereby *represent* those objects, is to fall prey to encodingism. Such correspondences, should they be definable, may be clear to us, as observers/users of the system, but how is the system itself supposed to know them? A theory of mind needs to *explain* how a system can know about the world, not simply *presuppose* that the system has this knowledge. The lack of a solution to this problem is precisely the empty symbol problem — the system can shuffle symbols endlessly, but these symbols remain contentless, ungrounded. As a hypothesis about the *internal workings of a computer*, the Physical Symbol System Hypothesis captures some important functional properties. As a hypothesis about *cognition*, however, the Physical Symbol System Hypothesis is fatally deficient.

And it is flawed precisely because of its commitment to encodingism. Given our argument thus far, this should be no surprise; making the point with respect to a well-known project in Artificial Intelligence, however, illustrates concretely the pervasiveness of encodingism in AI.

Newell bounces between the horns of the computer-versus-cognition dilemma. He clearly is most interested in (and on the safest ground!) viewing symbols solely as internal states.

> The primitive symbolic capabilities are defined on the symbolic processing system itself, not on any external processing or behaving system. The prototype symbolic relation is that of access from a symbol to an expression *[i.e. another internal object]*, not that of naming an external object. (Newell, 1980a, p. 169)

However, he occasionally states that, even though it is of secondary importance, symbols can correspond to objects in the world.

> Then, for any entity (whether in the external world or in the memory), ... processes can exist in the

> symbol system ... that behave in a way dependent on the entity. (Newell, 1980a, pp. 156-157)
>
> ... the appropriate designatory relations can be obtained to external objects ... (Newell, 1980a, p. 169)

Our central critique of the Physical Symbol System Hypothesis, then, is that it focuses on the processing of internal indicators or "symbols" while giving no answer whatsoever to how these "symbols" can have representational content. As a framework for understanding cognition, this absence is fatal. Four additional points only darken the cloud of confusion.

First, Newell's notion of *designation* (which he later extends to representation) is so general as to be vacuous. Nevertheless, this is the ground — and, therefore, the limit — of Newell's attempt to address epistemic contact and content.

> *Designation*: An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y. (1980a, p. 156)

This definition permits descriptions such as "a transmitter molecule docking on a cell receptor designates, relative to internal processes of the cell, the activities of the preceding neuron that released the transmitter" and "the octane of the gasoline put into the car's tank designates, relative to the internal processes of the engine, the octane of the gasoline in the underground tank from which it was filled" and "the key strokes on my keyboard designate, relative to the internal processes of the computer, the intentions and meanings that I am typing." These all involve various kinds of correlational and functional or control relationships, but *none* of them are representational relationships. This "model" is an impoverished "correspondence plus subsequent appropriate function" notion of encodingism. It is impoverished in the sense that the core of the entire definition is in the word "depends," but, as shown elsewhere (Bickhard, 1993a) and below, it is fundamentally inadequate even if that functionality is elaborated, even if "depends" is explicated (e.g., Smith, 1987).

Second, Newell has a deficient notion of a system being embedded in, and interacting with, its environment. **Input** and **behave** are just two not very important functions of a physical symbol system; there is no sense of the representational importance of *inter*action with an

environment. And he has no notion whatsoever of the constitutive role of goal-directed interaction in representation.

Third, note that, although designation, and therefore "representation," are transitive, genuine representation is *not* transitive. If X represents Y — e.g., X is a name for Y — and Y represents Z — e.g., Y is a map of Z — it does not follow that X represents Z. You could not find your way around merely by having the name for the appropriate map. This divergence with respect to transitivity is a clear difference between informational — correspondence, tracking, isomorph, and so on — relationships, and the possibility for control relationships that they provide, which are transitive, and true representational relationships, which are not transitive.

Finally, Newell mentions briefly that processing a symbolic representation can result in an "unbounded" number of new representations (1980a, p. 177). This is true, in the sense that applying a finite set of operators to a finite set of basic elements can result in an infinite set of non-basic elements. However, this process cannot result in fundamentally new representations. The infinite set of integers can be derived by applying one operator (successor) to one basic element (zero). Nevertheless, there is no way to derive the real numbers (nor anything else) from this set of basic elements and basic operations. If what is needed is not in the set, it does not matter that the set might be infinite. For example, the space of *even integers* is infinite, but that doesn't help much if you need an odd integer — or rational or real or complex or quaternion or matrix or tensor or fibre bundle connection — or representation of a car or a steak — or democracy or virtue — and so on.

The Physical Symbol System Hypothesis, then, at best captures some of the internal and external functional relationships that might exist in a computer, but it does not genuinely address *any* of the issues of representations. It can be construed representationally only by stretching the internal pointer relationships in and among data structures to an analogous notion of pointing to things in the world. But what is being "pointed to" in a computer is hardwired to be functionally accessible (and even then is *accessed*, not represented), and this has nothing to do with representation of the external world. On the alternative sense of access, correspondences simply do not constitute representations, no matter how useful they may be for various sorts of control relationships and the consequent functional sensitivities that they can provide. We are in strong agreement with the goal of naturalizing representation that is

inherent in the very notion of a physical symbol system, but this hypothesis has not achieved that goal.

The Problem Space Hypothesis

Overview. In addition to the framework of the Physical Symbol System Hypothesis, SOAR is based on the secondary *Problem Space* hypothesis (Newell, 1980b). This is the hypothesis that all symbolic cognitive activity can be modeled as heuristic search in a symbolic problem space. In particular, Newell claims that reasoning, problem solving, and decision making can all be captured as searches in appropriately defined problem spaces.

A problem space is a set of encoded states interconnected by possible transformations. The space is usually an implicit space defined by the combinatoric possibilities of some set of basic encodings, and the transformations are similarly atomized and encoded. In this space, an initial state and a goal state are specified and the abstract task is to find a path from the initial state to the goal state via the allowed transformations. The general problem space model is supposed to capture variations across reasoning, problem solving, and decision making with corresponding variations in what the state encodings and the transformational encodings are taken to encode. There is a clear and fundamental dependency on the Physical Symbol System Hypothesis here, with its fatal presupposition of encodingism.

Critique. We find here a more subtle and even more serious consequence of the encodingist presupposition, however. The problem space hypothesis can be construed, in a minimal information form, as a trial and error search in a space of possibilities defined by the combinatoric space of some generating set of explicit atomic encodings. The criterion for the search is some structure of encodings that satisfies the goal definition. In more than minimal information cases, the search need not be blind trial and error, but can use heuristic information to enhance the process; it might even become algorithmic. But this not only presupposes encodingism in the presumed implementation of the problem space, it inherently restricts all such variation and selection searches to the combinatoric possibilities given by the generating set of atomic encodings.

In particular, there is no possibility in this view of generating new *emergent* representations as trials toward possible solution of the problem, as possible satisfiers of the goal criteria. The only

"representational" states allowed are the syntactic combinations of already available atomic "representations." Put another way, the atomic encodings with which the system begins *must* be *already* adequate in order for the "cognitive" activity to possibly succeed, since no new representations outside of that combinatoric space are possible.

Newell, here, is committed to Fodor's necessary innateness ("pregivenness") of all basic concepts, with all of its bizarre consequences: inherent innate restriction on human cognitive capacities, innate but "non-triggered" representations for quarks and tensors and the U.S. Senate in the Neanderthal (since there is no way for evolution to have inserted those concepts since then), and so on (Bickhard, 1991b, 1991c; Piattelli-Palmarini, 1980). As pointed out earlier, Fodor's position is a massive *reductio* of the assumptions which Newell is presupposing (Bickhard, 1991a, 1991b, 1991c).

From a practical perspective, this means that the user of any hypothesis-space program must create *all* the necessary atomic encodings and must correctly *anticipate* which ones will be necessary in order for the system to work. Put still another way, the **construction of emergent representations** is one example of a cognitive process that cannot be modeled within the problem space hypothesis. Furthermore, historical problem solving — in physics or mathematics or ethics — does involve the creation of new representations — representations not anticipated in the simple combinatorics of previous representations. Clearly, in this fundamental sense at least, the Problem Space Hypothesis is not adequate to model genuine intelligence.

In fact, most problem solving does not involve pregiven spaces of possible states and solutions: problem spaces. The construction of appropriate possible solutions — which may involve the construction of emergent representations, and may or may not involve organizations of such "state" possibilities as problem spaces — can often comprise the most difficult part of problem solving — or reasoning, or decision making. Historical examples can even involve rational reformulations of what is to *count as* a solution — rational reformulations of problem definitions (Nickles, 1980). Even in relatively trivial problems, such as missionary and cannibals problems, the generation of new elements and attributes for the basic state language, the generation of appropriate "action" representations, and *theorem finding* — not just theorem proving — concerning properties of the problem and the "problem space" can all be critical in effective and tractable problem solving (Amarel, 1981). The

problem space hypothesis is, *in-principle*, incapable of capturing such cognitive phenomena.

SOAR

Overview. We turn now to the SOAR project (Laird, Newell, & Rosenbloom, 1986). The goal of the SOAR project is to define an architecture for a system that is capable of general intelligence. SOAR explicitly follows the Physical Symbol System Hypothesis, so it illustrates nicely the practical consequences of encodingism. As a "cognitive" system, SOAR is wholly a model of internal processing for a system, and needs a programmer/user to do all the representational work for it.

SOAR is fundamentally a search architecture. Its knowledge is organized around tasks, which it represents in terms of problem-spaces, states, goals, and operators. SOAR provides a problem-solving scheme — the means to transform initial states of a problem into goal states. One of the major advances SOAR claims is that any (sub-)decision can be the object of its own problem-solving process. For example, if SOAR is attempting to play chess and does not know which move to make in a certain situation, it can pose the problem "choose which move to make" to itself; work on this in a new, subordinate problem-space; then use the result to decide what move to make in the original space. This property is referred to as *universal sub-goaling*.

Another claimed advance is the ability to combine sequences of transformations into single *chunks.* In SOAR, this is a richer process than just the composition of the component transformations. It allows, for example, for a form of generalization of the conditions under which the chunked transformation is to be applied. The process, of course, is referred to as *chunking*.

As should be clear, SOAR is a model of internal processing for symbol manipulation systems. Laird, Newell, & Rosenbloom are explicit about their user/programmer version of encodingism, stating that SOAR "encodes its knowledge of the task environment in symbolic structures." However, to be precise, it is not SOAR that does the actual encoding. Programmers do the actual representational work of encoding a problem in terms of states, goals, operators, and even evaluation metrics.

Critique. Thus, SOAR already can be seen to be just another example of an encodingist Artificial Intelligence system. However, since SOAR is well-known and influential, it is worth considering in a bit more

detail how encodingism subverts the worthwhile goals of the project. We'll do this by considering how several interrelated aspects of SOAR that the authors take to be very important — universal sub-goaling and chunking — are weakened by SOAR's programmer-specified semantics.

Universal Sub-goaling. Laird, Rosenbloom, and Newell consider universal sub-goaling, the property of being able to do problem solving to make any decision, to be one of the most important contributions of SOAR. An example they discuss in detail is taken from the 8-puzzle problem. Suppose that at a given point, SOAR does not know whether it is better to move a tile in the puzzle right, left, up, or down. It creates a goal of choosing between these four operators and sets up a problem space to solve the goal. There are two methods that SOAR can use to do search in this space.

- If it has a metric for evaluating the goodness of states, it can apply each of the operators, use the metric to evaluate the resulting states, and decide to use the operator that resulted in the highest valued state. However, this is only possible if SOAR's programmer has provided it with an evaluation metric.
- If it does not have a metric, SOAR will continue to recurse until it solves the problem. That is, it will apply the operators and come up with four states among which it cannot distinguish. It will then set up the problem of deciding which of these states is best. It will continue on until it reaches the goal state.

That is, if SOAR's programmer has provided it with an evaluation metric, SOAR will use it, and, if not, SOAR will do a depth-first search. The flexibility of being able to use whatever evaluation metric a programmer provides is a convenient modularization of its search process, but it is not more than that. The ability to iterate its process of setting up (sub-)goals with associated problem spaces and evaluation metrics etc. — so long as all the necessary encoding for all those problem spaces and metrics has already been anticipated and provided by the programmer (such encoding frameworks can sometimes be reused at higher levels of recursion) — is, again, a convenient re-entrant modularization of the search process, but it is not more than that. And it is not even particularly convenient, given that *all* relevant information must be anticipated and pregiven by the programmer.

One example of this is SOAR's "learning" of a "macro-operator" solution to the eight-puzzle problem (Laird, Rosenbloom, Newell, 1986; Laird, Newell, Rosenbloom, 1987). These macro-operators constitute a *serial decomposition* of the general problem, where a serial decomposition is one in which the attainment of each successive subgoal leaves all previous subgoals intact. In this case, the successive goals have the form: 1) place the blank in the proper location, 2) place the blank and the first tile in the proper locations, 3) place the blank and the first two tiles in the proper locations, and so on. On the one hand, SOAR's ability to develop this macro-operator solution is deemed to be of "particular interest" because SOAR is a general problem solver and learner, rather than being designed specifically for the implementation of macro-operators (Laird, Rosenbloom, Newell, 1986, p. 32). On the other hand, in order for SOAR to accomplish this feat, it must be fed *two* complete problem spaces — one defining the basic eight puzzle and one defining the macro-operator version of the eight puzzle (Laird, Rosenbloom, Newell, 1986; Laird, Newell, Rosenbloom, 1987). Further, it must be hand tutored even in order to learn all the macro-operators, once fed their definitions (Laird, Rosenbloom, Newell, 1986, p. 37) (though this is probably a matter of speed of computation). Still further, the macro-operator characterization of the eight-puzzle is itself due to Korf (1985) (or predecessors), so this is an example of a historically and humanly developed problem space characterization — not one developed by SOAR or by any other program. In sum, SOAR can accomplish a serial decomposition of the eight-puzzle problem *if* it is fed a basic eight-puzzle problem space and if it is fed a macro-operator space capturing that serial decomposition that someone else has already figured out. This is an enormous collective labor for an "accomplishment" that is in fact rather boring. Truly, SOAR programmer(s) must do *all* of the hard work.

The claims made for Universal Subgoaling, however, are extreme indeed. It is claimed, for example, that "SOAR can reflect on its own problem solving behavior, and do this to arbitrary levels" (Laird, Newell, Rosenbloom, 1987, p. 7), that "Any decision can be an object of goal-oriented attention." (Laird, Newell, Rosenbloom, 1987, p. 58), that "a subgoal not only represents a subtask to be performed, but it also represents an introspective act that allows unlimited amounts of meta-level problem-space processing to be performed." (Rosenbloom, Laird, Newell, McCarl, 1991, p. 298), and that "We have also analyzed SOAR in terms of concepts such as meta-levels, introspection and reflection"

(Steier et al, 1987, p. 307). It would appear that SOAR has solved the problem of conscious reflection. However, in Rosenbloom, Laird, and Newell (1988) it is acknowledged that what is involved in SOAR is a *control* notion of *recursiveness*, not an *autoepistemic* notion such as "quotation, designation, aboutness, or meta-knowledge" (p. 228). Such recursiveness, with perhaps some more convenient than hitherto modularizations of the recursive processes, is in fact all that is involved in universal sub-goaling. SOAR's claims to such phenomena as "reflection," "attention," and "introspection," then, are flagrantly bad metaphorical excesses made "honest" by redefinitions of the terms (into a "control" tradition) in a secondary source paper (Steier at al, 1987; Rosenbloom, Laird, Newell, 1988).

Chunking. The second major innovation in SOAR is the process of Chunking (Laird, Rosenbloom, Newell, 1984, 1986; Laird, Newell, Rosenbloom, 1987; Steier et al, 1987; Rosenbloom, Laird, Newell, McCarl, 1991). Chunking is supposed to constitute a "general learning mechanism." Together with universal subgoaling, then, SOAR has supposedly solved two of the deepest mysteries of the mind — consciousness and learning. As might be expected, however, there is less here than is first presented to the eye.

Chunking is to a first approximation nothing more than the composition of sequences of productions, and the caching of those resultant compositions. When this works well, appropriate initial conditions will invoke the cached composition as a unit, and save the search time that was involved in the construction of the original sequence of productions. This is useful, but it clearly does not create anything new — it saves time for what would have ultimately happened anyway. No new representations are created, and no hitherto unconstructable *organizations* of encodings arise either. Composition of productions is fundamentally inadequate (Neches, Langley, Klahr, 1987).

Chunking's claim to fame, however, does not rest on production rule compositionality alone. In addition, chunking permits generalization in the conditions to which the compositions can apply. Such generalization occurs in two ways. First, generalization occurs via variabilization — the replacement of identifiers with variables. This makes SOAR "respond identically to any objects with the same description" (Laird, Newell, Rosenbloom, 1987, p. 55). And second, generalization occurs via "implicit generalization" which functions by "ignoring everything about a situation except what has been determined at

chunk-creation time to be relevant. ... If the conditions of a chunk do not test for a given aspect of a situation, then the chunk will ignore whatever that aspect might be in some new situation." (Laird, Newell, Rosenbloom, 1987, p. 55).

Both variabilization and implicit generalization are forms of ignoring details and thereby generalizing over the possible variations in those details. This can be a powerful technique, and it is interesting to see what SOAR does with it. But, only identifiers already created by the programmer in slots already created by the programmer can be "ignored" (variabilized), and only aspects of situations (slots) already created by the programmer can be disregarded, and, thus implicitly generalized over. In other words, chunking functions by eliminating — ignoring — encodings and encoding slots that are programmer pregiven. Again, nothing new can be created this way, and the generalizations that are possible are completely dependent on the encoding framework that the programmer has supplied.

This dependence of SOAR's "learning" on preprogrammed encoding frameworks holds in two basic senses: 1) There is a nearby outer limit on what can be accomplished with such elimination generalizations — when everything pregiven has been eliminated, nothing more can be eliminated (generalized over). 2) The generalizations that *are* available to such elimination methods are completely determined by those preprogrammed encoding frameworks. In other words, an aspect can be generalized over only if that aspect has already been explicitly pre-encoded, otherwise there is nothing appropriate to ignore and thus generalize over. This latter constraint on SOAR's generalization abilities is dubbed "bias": "The object representation defines a language for the implicit generalization process, bounding the potential generality of the chunks that can be learned" (Laird, Rosenbloom, Newell, 1986, p. 31).

Just as the programmer must anticipate all potentially relevant objects, features, relationships, atomic actions, etc. to be encoded in the problem space in order to make SOAR function, so also must the programmer anticipate the proper aspects, features, etc. that it might be relevant to ignore or variabilize, and, thus, generalize over. As a form of genuine learning, chunking is extremely weak. From a representational perspective, the programmer does ***all*** the work. To construe this as a "general learning mechanism" is egregious.

Thus, not only is composition per se inadequate, composition *plus* "generalization" *plus* "discrimination" (the *addition* of encoded

constraints) are *collectively* incompetent, for example, for unanticipated reorganizations of encodings, reorganizations of processes, and the construction of new goals (Neches, Langley, Klahr, 1987; Campbell & Bickhard, 1992b). The SOAR architecture, and, ipso facto, any implementation of that architecture, does not escape these failures.

Summary Analysis. SOAR is far from the "architecture for general intelligence" it was touted to be. It cannot generate new representations, so it therefore cannot learn anything that requires representations not already combinatorically anticipated, nor decide anything, nor reason in any way that requires representations not already combinatorically anticipated (e.g., Rosenbloom, Newell, Laird, 1991). Among other consequences, it cannot recurse its problem spaces any further than has been explicitly made available by the programmer's encodings, despite the phrase "universal subgoaling." It cannot "reflect," despite the characterization of subgoal recursion as "reflecting." It cannot generalize in its chunking in any way not already combinatorically anticipated in the user provided encoding scheme for the problem space. SOAR is interesting for some the new possibilities within classical frameworks that it exemplifies and explores, but it cannot manifest any of the capabilities that are suggested by the terms used — "general intelligence," "reflection," "universal weak method learning," "generalization," and so on. In this respect, it is, at best, a massive example of "natural stupidity" (McDermott, 1981).

The multiple deficiencies of SOAR are not entirely unknown to SOAR's proponents. They are acknowledged in occasional brief passages that are inconsistent with such claims as "general intelligence," "reflection," "general learning," and so on. The deficiencies, however, are invariably treated as if they were mere technical problems, to be conclusively fixed and solved in future elaborations of the system: SOAR

> can not yet learn new problem spaces or new representations, nor can it yet make use of the wide variety of potential knowledge sources, such as examples or analogous problems. Our approach to all of these insufficiencies will be to look to the problem solving. Goals will have to occur in which new problem spaces and representations are developed, and in which different types of knowledge can be used. The knowledge can then be captured by chunking.
> (Laird, Rosenbloom, Newell, 1986, p. 43).

Not only is the language in which SOAR is presented flagrantly overblown (making claims for SOAR that SOAR has not even touched) but this "faith in principle" in the general approach ("all problems will succumb to more of the same") is the most basic disease of invalid research programmes. SOAR is inherently an instance of the problem space hypothesis, and, a fortiori, of the Physical Symbol System Hypothesis (Norman, 1991). Each of these, in turn, inherently presupposes encodings as the fundamental nature of representation, which entails the impossibility of the emergence of new representation out of non-representational phenomena. But, until genuinely emergent representation is possible (among other things), neither genuine intelligence, nor reasoning, nor problem solving, nor decision making, nor learning, nor reflection will be possible. Any gestures that SOAR might make in these directions will have to be already effectively anticipated in the programmer supplied encodings.

Problem spaces (necessarily pregiven) for the construction of problem spaces might conceivably have some practical value in some instances, but such a notion merely obfuscates the fundamental in-principle issues. Either an encoding framework can successfully anticipate all possibly needed representations or it cannot. The incoherence argument, and related arguments, show that it cannot. And, therefore, since SOAR fundamentally exemplifies the encodingist approach, it is impossible for it or anything within its framework to make good on its claimed aspirations.

Furthermore, and most importantly, the restrictions and impossibilities that encodingism imposes on SOAR and on the problem space hypothesis more generally are simply instances of the restrictions and impossibilities that encodingism imposes on all of Artificial Intelligence and Cognitive Science. The physical symbol system model is simply one statement of the encodingism that pervades and undergirds the field. And it is a fatally flawed foundation.

PROLIFERATION OF BASIC ENCODINGS

Any encodingism yields an ad hoc proliferation of basic encodings because of the impossibility of accounting for new kinds of representation within the combinatoric space of old basic encodings. Encodingism cannot account for the emergence of new representational content; it can only account for new combinations of old contents. The incoherence problem turns precisely on this impossibility of encodingism to be able to

account for new, foundational or basic, representational content. Because the emergence of new sorts of encoding elements is impossible, *any* new representational content requires an ad hoc designed new element to represent it. In relatively undeveloped programmatic proposals, this difficulty can be overlooked and obscured by simply giving a few examples that convey the appearance of being able to reduce representation to combinations of elements — e.g., the famous case of the "bachelor = unmarried male," or the semantic features proposal for language (Katz & Fodor, 1971; Chomsky, 1965). Whenever such a programme is taken seriously, however, and a real attempt is made to develop it, the impossibility of capturing general representation in an encoding space makes itself felt in a proliferation of elements as more and more sorts of representational contents are found to be essential that cannot be rendered as combinations of those already available (e.g., Bolinger, 1967).

CYC — Lenat's Encyclopedia Project

Doug Lenat and his colleagues at the Microelectronics and Computer Technology Corporation (MCC) are engaged in a project that directly encounters this problem of the proliferation of basic encodings (Lenat & Guha, 1988; Lenat, Guha, & Wallace, 1988; Guha & Lenat, 1988). They are attempting to construct a massive knowledge base containing millions of encoded facts, categories, relations, and so on, with the intent that the finished knowledge base will define our consensus reality — will capture the basic knowledge required to comprehend, for example, a desk top encyclopedia. This effort is the en**CYC**lopedia project.

It's All Just Scale Problems. Lenat and colleagues are well aware of the tendency for knowledge bases, no matter how adequate for their initial narrow domains of knowledge, to be fundamentally not just incomplete, but inappropriate and wrongly designed in attempts to broaden the knowledge domain or to combine it with some other domain: Categories and relations are missing; categories are overlapping and inconsistent; categories and relations that need to have already been taken into account, even in a narrow knowledge base, were not taken into account because the distinctions weren't needed so long as that narrow domain was the limit of consideration; the design principles of the knowledge base are inadequate to accommodate the new domain contents and relationships; and so on. Knowledge bases do not scale up well.

The suggestion in Lenat's project is that these problems — representational proliferation, representational inconsistency and redundancy, design inappropriateness, and so on — are *just* scale problems, and, therefore, will be overcome if the scale is simply large enough to start with. The suggestion is given analogical force with an onion analogy: concepts and metaphors are based on more fundamental concepts and metaphors, which are based on still more fundamental ones, like the layers of an onion, and, like an onion, there will be a central core of concepts and metaphors upon which all else are based. These central notions, therefore, will be adequate to *any* knowledge domain, and, once they are discovered, the scale problem will be overcome and the proliferation problem will disappear. All new concepts will be syntactically derivable from concepts already available, and, ultimately, from the basic "onion core" concepts. The "onion core," then, is supposed to provide the semantic primitives adequate to the construction of everything else (Brachman, 1979).

The Onion is not an Argument. There are at least three problems with the position. The first is that Lenat et al give *no argument whatsoever* that this will be the case or should in any way be expected to be the case. The onion analogy is the *only* support given to the hoped for convergence of needed concepts — a convergence in the sense that, after a large enough base has been achieved (literally millions of facts, categories, etc., they say), the core of the onion will have been reached, and, therefore, concepts and relations etc. needed for *new* material to be incorporated into the base will already be available. The entire project is founded on an unsupported onion analogy.

There is, in fact, a puzzle as to why this would seem plausible to *anyone*. We venture the hypothesis that it is because of the intuition that "encodings are all there is" and a similar intuition from the innatists that people are born with a basic stock of representational raw material.

The Onion Core is Incoherent. The second problem is that the presumed core of the representational onion is "simply" the base of *logically independent grounding encodings*, and the circular incoherence of that notion insures that such an encodingist core cannot exist. From a converse perspective, we note that the layered onion analogy *is* appropriate to the purely syntactic combinatorialism of encodingism, but that the invalidation of encodingism *ipso facto* invalidates any such combinatorically layered model of the organization of representation in general.

Combinatorics are Inadequate. The invalidity of the presumed combinatoric organization of possible representations, in turn, yields the third problem: the supposed combinatoric scale problem proves impossible to solve after all. It is *not* merely a scale problem. New concepts are rarely, if ever, simply combinations of already available encodings, and, therefore, cannot in principle be accommodated in a combinatoric encoding space — no matter how large the generating set of basic encodings. New representation is a matter of emergence, not just syntactic combination, no matter what the scale might be. The space of possible representations is *not* organized like an onion.

Note that this is an in-principle impossibility. Therefore, it is not affected by any issues of the sophistication or complexity of the methods or principles of such syntactic combinatorialism. That is, the various fancy apparatuses of exceptions, prototypes, default logic, frame systems with overrideable defaults — however powerful and practically useful they may be in appropriate circumstances — do not even address the basic in-principle problem, and offer no hope whatsoever of solving it.

There are several different perspectives on the intrinsic inadequacy of combinatorial encoding spaces. We take this opportunity of the discussion of the CYC Project to discuss four of them, of successively increasing abstraction. The fourth of these perspectives involves technical arguments within logic and mathematics. Some readers may wish to skip or to skim this section (Productivity, Not Combinatorics).

Ad hoc Proliferation. First we must point out again the history of the ad hoc proliferation of encoding types in every significant attempt to construct an encodingism. Both internal to particular projects, such as feature semantics, as well as in terms of the historical development from one project to another, new kinds of encoding elements have had to be invented for every new sort of representational content. In fact, the practical power and realistic applicability of encodingist combinatorialism has proven to be extremely limited (Bickhard, 1980b, 1987; Bickhard, & Richie, 1983; Bolinger 1967; Fodor, 1975, 1983; Shanon, 1987, 1988, 1993; Winograd & Flores, 1986). This, of course, is precisely the history that Lenat et al note, and that they claim — without foundation — is merely a scale problem.

Historically False. A second perspective on the inadequacy of combinatorialism is a historical one. In particular, an encounter with the necessity of the proliferation of new sorts of representations can be found

in any history of any kind of ideas. New ideas are not just combinations of old ideas, and any such history, therefore, comprises a massive counterexample to any combinatorialism of concepts — and, therefore, to encodingism. Lenat's knowledge base, more particularly, could not capture the history of mathematics within its onion unless that history were already included in, designed into, the system. This is exactly the point that he brings against all previous knowledge base projects, and the point that is supposed to be overcome because this project is Big. Even more to the point, Lenat's onion will in no way *anticipate* the future history of mathematics, or any other field of ideas, in its generated combinatoric space.

Another perspective on this point is provided by the realization that encodings cannot capture generalization, nor differentiation, except in terms of the encoding atoms that are already available in the encoding space. Abstraction as a reduction of features, for example, can only proceed so long as the atomic features are already present and sufficient. Differentiation as an intersection of categories, for another, similarly can only proceed in terms of the (sub)categories already encoded. These are just special cases of the fact that encodings cannot generate new representations, only at best new combinations of representations already available.

The history of mathematics, to return to that example, is a history of deep, complex, and often deliberate abstractions from earlier mathematics and from other experience (MacLane, 1986). No combinatoric onion can capture that. To posit that any atomic rendering of Babylonian mathematics would be combinatorically adequate to contemporary mathematics is merely absurd. Why would anyone think that an atomic rendering of *today's* mathematics, *or any other domain*, would fare any better in our future? The point remains exactly the same if we shift to Babylonian and contemporary culture writ large: Babylonian culture would have to have contained all contemporary culture (including contemporary mathematics) in the combinatorial space of its encoding representations.

Furthermore, mathematical abstraction is often an abstraction of relations, not of objects or predicates. The relational structures that define groups or fields or vector spaces or lattices (Birkhoff, 1967; Herstein, 1964; MacLane, 1971, 1986; MacLane & Birkhoff, 1967; Rosenfeld, 1968) would be easy examples. Relational encodings cannot be constructed out of element and predicate encodings (Olson, 1987).

Therefore, Babylonian mathematics could be combinatorically adequate to modern mathematics only if those critical relational encodings (set theory, category theory, group and field theory per se?) were *already present* in Babylonian (prehistoric, prehuman, premammal, prenotochord?) times. Modern relational concepts could then be "abstracted" by peeling away whatever festooning additional encodings were attached in earlier times, leaving only the critical relational encodings for the construction of modern conceptions.

Clearly, we are in the realm of a Fodorian radical innatism of everything (Fodor, 1981b).[5] But "the argument has to be wrong, ... *a nativism pushed to that point becomes unsupportable, ... something important must have been left aside*. What I think it shows is really not so much an a priori argument for nativism as that *there must be some notion of learning that is so incredibly different from the one we have imagined* that we don't even know what it would be like as things now stand" (Fodor in Piattelli-Palmarini, 1980, p. 269). What is in error about current conceptions of learning is that they are based on false conceptions of representation — encoding conceptions. Encoding models of representation force a radical innatism, and Lenat is just as logically committed to such an innatism as any other encodingist (Bickhard, 1991c, 1993a). Lenat's onion-core would have to anticipate the entire universe of possible representations.

The Truth Predicate is not Combinatorial in L. The third perspective on the inadequacy of syntactic combinatorialism is a counterexample from Tarski's theorems regarding Truth predicates, as discussed earlier. In particular, any language that is "adequate to its own semantics" is a language in which that language's own Truth predicate can be constructed, and any language which can contain its own Truth predicate is logically inconsistent. An inconsistent language, in turn, cannot contain *any* coherent capturing of its own semantics, since any statements of semantic relationships can be validly (within the inconsistent language) replaced by their negations. Syntactic combinatorialism is limited to constructions within a given encoding

[5] When did those encodings get inserted into the genes? And how could they have been inserted? Where did they come from? If they were somehow emergently constructed by evolution, then why is it impossible for learning and development in human beings to emergently — non-combinatorically, non-onion-like — construct them? If they cannot have been emergently constructed, then evolution is just as helpless as learning and development, and representations become impossible: they did not exist at the Big Bang, so they could not have emerged since then (Bickhard, 1991c, 1993a).

language, and, by these theorems, syntactic combinatorialism is intrinsically incapable of consistently, coherently, defining Truth for that encoding system. The Truth predicate itself, then, for any encoding language, is a straightforward counterexample to any purported adequacy of syntactic combinatorialism to be able to capture the space of possible representations. It is simply impossible.

Productivity, not Combinatorics. Our fourth perspective on the intrinsic inadequacy of any combinatoric encoding space is an abstract in-principle mathematical consideration. Any combinatoric encoding space will be (recursive, and, therefore) recursively enumerable. The set of possible principles of functional selection, on the other hand, and, therefore, of interactive functional representation, will be at least productive (Rogers, 1967; Cutland, 1980).

A productive set is a set **S** for which there exists a recursive function **F** such that for any recursively enumerable $\mathbf{S_1}$ contained in **S** with index **x**, **F(x)** will be an element in **S** but not in $\mathbf{S_1}$. That is, any attempt to capture a productive set by recursive enumeration yields a recursively computable *counterexample* to the purported recursive enumeration. The True well-formed formulas of elementary arithmetic are productive in this sense: any purported recursive enumeration of them will recursively yield a counter-example to that purported recursive enumeration.

The basic realization involved here is that interactive representation is intrinsically functional, not atomistic. Any encoding or encoding combination can do no more than influence *functional selections* in the ongoing process of a system, but the space of possible such functional selections *is* the space of possible interactive representations, and that space is generated as possible *functional organizations that might be selected*, *not* as possible combinations of elements of some finite set of atomic possible selections. New *kinds* of selections, thus new kinds of representations, can occur given new *kinds* of functional organizations. There are no atomic representations in this view.

Conversely, a counterexample can be constructed for any given purported encoding enumeration by constructing a new functional organization, and, thus, a new possible representational selection, that differs internally or in some other intrinsic sense from the functional organizations that are selected for by all available atomic “encodings.”

In fact, the existence or definability of *any* productive set constitutes a counterexample to *any* programme of atomic combinatorialism, since these sets are not non-circularly definable nor constructable as mere combinatorialisms from some, or any, atomic base set — if they were so definable or constructable, they would not be productive since they would then be capturable by a recursive enumeration. The very *existence* of productive sets, then, demonstrates that the space of possible forms and patterns of representation-as-functional-selection cannot be captured atomistically, combinatorically, and, therefore, cannot be captured within any encodingism. Productive sets cannot be non-circularly defined explicitly, syntactically, on any atomic base set; they can, however, be defined implicitly.[6]

Any recursive enumeration (encoding model) *within* a productive set **S** (space of possible interactive functional representations) yields its own recursively generable *element of **S*** (new interactive representation) that is *not included in the enumeration* (not included in the encoding space). The enumeration, therefore, is not complete. That new element can be included in a *new* recursive enumeration (e.g., defined as a new atomic element of the generative encodings), which will generate its own exception to that new enumeration (encoding system) in turn. This can yield still another exception to the enumeration, which could be included in still another enumeration, and so on. In other words, it is impossible to capture a productive set by a recursive enumeration, and any attempt to do so embarks on this proliferative unbounded expansion of attempting to capture counterexamples to the last attempted enumeration — the ad hoc proliferation of encoding types. The enumeration (encoding system) *cannot* be complete. This futile pursuit of a productive set with an enumeration is the correct model for the relationship between representation and encodingisms. It is far different than, and has opposite implications from, Lenat's onion metaphor.

Isn't Infinite Enough? One prima facie rejoinder to this point would be to claim that, although encodingism suffers from a combinatoric limitation, nevertheless an encoding combinatoric space is infinite in extent, and that ought to be enough. Infinite it may be, but if it does not contain the correct representations, the ones needed for a given task, that does no good. The set of even integers is infinite, but that is of no help if what is needed is an odd integer, or real, or a color or a food, and so on.

---

[6] For related issues, see the discussion of the Frame Problems below.

The possible interactions of a simple finite automaton can also be infinite, but that does not imply that finite automata theory suffices relative to Turing machine theory.

A Focus on Process Instead of Elements. A natural perspective on representation from within an encoding perspective is to focus on the set of possible combinations of basic encoding elements — on the set of possible encoding representations. This is the perspective in which the above considerations of formal inadequacies of encodingism are presented. A different perspective on representation and encodingism, one more compatible with interactivism, emphasizes the processes involved rather than the sets to which those processes are computationally competent — competent as enumerators or detectors, and so on. The *combinatoricism* of encodingism, as a form of *process*, is clearly drastically inadequate to the formal processes of Turing machine theory. The inadequacy of Turing machine theory, in turn, to be able to capture interactive representation provides still another perspective on the fundamental inadequacy of encodingism. That is, in the interactive view, the potentialities of *representation* are an *aspect* of the potentialities of (certain forms of) *process* — unlike the diremption of representation from process in the juxtaposition of Tarskian model theory and Turing machine theory, of semantics and computation. Furthermore, interactive representation is an aspect of a *form* of process that cannot be captured by Turing machine theory, and certainly not by any simple encodingist combinatorialism. The *process* weakness of encodingism, therefore, constitutes a *representational* inadequacy relative to interactive representation.

As a set, then, the free space of an encodingism is intrinsically too small, and, as a process, the combinatorialism of an encodingism is inherently too weak. Any attempt to capture representation in an encodingism, then, is doomed to the futile chase of ever more not-yet-included representational contents, is doomed to an inevitable proliferation of basic encoding elements in an attempt to capture representations not included in the prior space. An encoding space is always too small and the combinatoric process is always too weak to be adequate to all possible representations.

Slot "Metaphor" versus Genuine Metaphor. There are yet other problems with Lenat's project. One is that he proposes a model of metaphor as a mapping of slots to slots between frames. This is probably about as much as can be done within a slot-and-frame encoding model,

but why it should be taken to be adequate to the creativity of genuine metaphor is not clear and is not argued. Furthermore, the onion analogy itself is rendered in terms of metaphors built on metaphors, etc. down to a presumed core of metaphors. Whatever plausibility this might seem to have derives from the inherent constructive *creativity* of genuine metaphor: mappings of slots, on the other hand, require slots (and frames) that are *already present*. The onion, therefore, is not only an analogy in lieu of an argument, it is an analogy built on a fundamental equivocation between genuine metaphor and the impoverished notion of "metaphor" as slot to slot mappings.

This is an implicit circularity in Lenat's onion: the onion analogy is used to make the scope claims of Lenat's encodingism project seem plausible, but the onion model itself is made plausible only in terms of the layering of metaphors, and Lenat's encoding model for those metaphors, in turn, — slot-to-slot mappings between frames — presupposes the validity of the general encoding approach that the onion metaphor was supposed to make plausible in the first place. Lenat's onion is hard to swallow.

Contradiction: Does Pushing Tokens Around Suffice, or Not? Still another problem in this project is that, although there is much discussion of the fact that the semantics of knowledge bases are in the user, not in the system — a point we clearly agree with — there is later a discussion of the massive knowledge base in *this* project as if *it* would understand, would have its own semantics. This issue *is* addressed, though hardly in a satisfactory way. In fact, "Yes, all we're doing is pushing tokens around, but that's all that cognition is." (Lenat & Guha, 1988, p. 11) The basic claim is that somehow by moving to the massive scale of this knowledge base project, the tokens inherently acquire semantics for the system, not just for the user. As with the proliferation problem, sufficient scale is supposed to solve everything in itself. The magic by which this is supposed to happen is not specified.

Accountability? The only discernible consequence of these incantations of massive scale is that the project *cannot be accountable* for any of its claimed goals till sufficient scale (itself only vaguely specified) has been reached, which means, of course, until massive amounts of time and money have already been spent. Here, as elsewhere in the CYC Project documents, the glib and breezy style seems to have dazzled and confused not just the intended readers, but the authors as well.

Claim: Dis-embodied, Un-situated, Un-connected Intelligence. Lenat and Feigenbaum (1991) provide a somewhat more sober presentation of the general project and strategy. But, although the tone is more sober, the claims are not. [7] There is still the explicit assumption that scale of knowledge base is what is of ultimately fundamental importance. In fact, there is an explicit hypothesis that no as-yet-unknown control structure is required for intelligence (1991, p. 192). Interactive control structures would seem to constitute a counter example to that, but they explicitly reject the notion that representation requires epistemic systems that are embedded in their environments. Action and interaction are not epistemically important. In later usage, in fact, they employ the notion of control structure as being synonymous with inference procedure (1991, p. 233); they don't have in mind any sort of real connection with the world even here.

Contradiction: The Onion Core is Situated, Embedded, Practices — It is NOT Dis-embodied Tokens After All. There is still the claim that the layers of analogy and metaphor "bottom out," though the onion per se is absent. The claim, however, is still unsupported (1991, p. 201). Later, in reply to Smith's critique (1991), they offer Lakoff and Johnson (1980) (1991, p. 246) as support for the existence of such a core to a universal conceptual onion. To the extent that this powerful book could be taken as supportive of any such core, however, any such support is dependent *not only* on the genuinely creative sense of metaphor — not capturable in mappings among pre-created slots — but also on the "core" being the sensory-motor ***practices*** of situated, embedded, human beings. The "core" that is being offered here in lieu of support for an onion core of combinatorically adequate representational atoms is instead a "core" of practices and forms of action and interaction upon which higher level metaphors are and may be created; it is *not* a "core" of grounding context independent encodings. As in the case of the original onion, only with a slippery inattention to fundamentals is any superficial appearance of support or supportive argument presented.

---

[7] Lenat, Guha, Pittman, Pratt, & Shephard (1990), in contrast, claim only a chance at surpassing a "brittleness threshold" — e.g., an inability to handle novel conditions — in knowledge bases. The problem of brittleness is, in a familiar way, claimed to be a scale problem — thus its construal in terms of a threshold — with the CYC project as the attempt to surpass the scale threshold. Again there is little argument for such characterizations. Nevertheless, with no claims to be creating genuine cognition, this is a relatively cautious presentation of the project.

Butchering Piaget. Although it is a small point in itself, it is also worth pointing out that this inattention and carelessness with claims is manifested in a flagrantly bad construal of Piaget — e.g., of Piaget's stage theory (p. 203, 204; cf. Bickhard, 1982, 1988a, 1988b; Bickhard, Cooper, Mace, 1985; Campbell & Bickhard, 1986; Drescher, 1986, 1991; Kitchener, 1986; Piaget, 1954, 1971, 1977, 1985, 1987). Pragmatic breeziness covers a multitude of sins.

False Programmes are not Falsifiable. Lenat and Feigenbaum (1991, p. 204) claim an advantage of falsifiability for their general approach. They present a commonsensical "let's try it and if it's falsified, then we'll learn something from that too" approach (e.g., p. 211). Unfortunately, they're never clear about what would constitute falsification (and the time horizons for any possible such falsification keep getting pushed back; their 1991 article projects into the next century). More deeply, they seem unaware that research programmes *cannot* be empirically falsified, even though they may be quite false. If the interactive critique is correct, then this entire project is based on false *programmatic* presuppositions. But, granting that they might conclude somehow that CYC had been falsified, on what would they place the blame? How would they diagnose the error? Perhaps CYC needs to be still bigger, or have more kinds of slots? Only conceptual level critique can discover and diagnose programmatic failure, but they are quite skeptical and derisive about such "mysticism" and "metaphysical swamp[s]" (p. 244; also, e.g., pp. 236-237).

Inconsistency: Claim Foundational Advances — Reject Responsibility. Lenat and Feigenbaum claim that a completed CYC system will actually have concepts and know things about the world (many places; e.g., pp. 244, 247), and yet they also reject pursuing the very issue of how concepts relate to the world (pp. 236-237). It seems that that issue is just another part of the metaphysical swamp. But it can't be both ways: Lenat and Feigenbaum cannot consistently claim solutions to foundational problems, and yet reject foundational critique. A careless "whatever works" becomes irresponsible when it both makes basic claims and rejects the very domain of the critique of such basic claims. There are not only a multitude of sins here, but very serious ones too.

Smith's Critique. Smith (1991) presents a strong critique of Lenat's project that has a number of convergences with our own. He, too, notes the absence of argument and of serious consideration of the deep and long standing problems involved. He also notes the absence of any

system semantics, in spite of apparent claims and presumptions to the contrary. Smith offers a number of critiques that, in our view, turn on the naive encodingism of this project, including the exclusive focus on explicit rather than implicit representation, and the absence of contextual, use, agentive, action, situatedness, or embodiment considerations. We wish only to second these criticisms, and to point out that these considerations are intrinsic aspects of interactive representation that are inevitably ruptured in the move to encodingism. They can at best be tacked on to an encodingist model to make an incoherent and ataxic hybrid. Lenat doesn't even attempt the hybrid.

TRUTH-VALUED VERSUS NON-TRUTH-VALUED

One of the basic choices that must be made when designing a system within an encoding approach concerns a critical aspect of the nature of the encoding primitives and their syntax. Should they be of the sort that takes on truth values — encodings of propositions? Should they be taken to encode sub-truth-value contents — such as perceptual or semantic features, concepts or categories? Or should they be of both kinds? The distinction between these two kingdoms of encodings — truth valued versus non-truth valued — poses a problematic for both theory and philosophy since no mere collection or structure of non-truth-value bearing encodings will intrinsically emerge as a truth value bearing encoding, no matter how formally syntactically correct that structure of elements is — there is a representational gulf between them. There is an unresolved aporia about how to get from sub-truth value encodings to truth value encodings. It is, of course, not only unresolved, but also unresolvable, since encodings can never introduce new representational content, and an emergent declarative or assertive *force* that yields a truth value ***is*** a new representational content.

If an investigator is restrictively concerned with the presumably propositional phenomena of thought and language (e.g., Anderson, 1983), then Frege's option is available of treating sub-propositional encodings as encodings carrying intrinsically *incomplete* meanings — incomplete relative to propositions — rather than being full representations of sub-truth value properties, features, categories, etc. (Dummett, 1973). Non-declarative sentences, of course, are difficult to accommodate within such a dedicatedly declarative framework (Bickhard, 1980b; Bickhard & Campbell, 1992). Specifying and formalizing the forms of such declarative-sentence incompleteness is the basic intuition that yields

categorial grammars (Bickhard & Campbell, 1992). This restriction to *exclusively* truth-valued encodings works only, however, so long as the problem of *origins* — logical, developmental, or evolutionary — of the presumed full propositional encodings is never addressed. That is, it *avoids* the problem of how to get propositions out of non-propositions — by presupposing truth valued encodings as its fundamental forms — but it does not address or solve that problem. That problem immediately encounters the gulf mentioned above, which cannot be crossed because of the impossibility of emergent representation within encodingism.

On the other hand, beginning with *non*-truth bearing encodings clearly does not solve any basic problems either. Not only do they encounter the incoherence problems directly just as much as do propositional encodings, but they simply halt on the other side from propositions of the gulf-of-impossible-emergence. An ad hoc postulation of *both* sorts of encodings is, of course, possible, but the problem of the gulf of emergence still remains, and is still unbridgeable.

In effect, this is "just" a particularization of the impossibility of emergence of encoding representation. On a ground of non-truth valued encodings, it is impossible to generate truth valued encodings; on a ground of truth valued encodings, it is impossible to generate non-truth valued encodings. The required emergence is impossible in either direction, since encoding emergence in general is impossible. The emergence *is required*, however, not only to get any representations on either side of the gulf in the first place, but also to bridge the gulf. The distinction between truth valued and non-truth valued, then, is another instance — a ubiquitous instance, a kind of *meta*-instance, since there will be many, unbounded, numbers of types of encoding elements *within* each side of the distinction — of the necessity of the ad hoc introduction of new types of encoding elements for new types of representation.

Procedural versus Declarative Representation

One area in which the general problem of the relationship between truth-valued and non-truth-valued encodings shows up is in the procedural-declarative "controversy" (e.g., Winograd, 1975). The general notion supposedly concerns a functional trade-off between declarative encodings and procedural encodings. In particular, the question arises whether declarative encodings, truth-valued-encodings, are in principle or in practice dispensable in favor of procedural encodings. Even overlooking that the procedures in this debate are themselves taken to be

encodings (e.g., programs, production rules, and so on), the encodingism representational framework so permeates the presuppositions of the discussions that what are considered to be purely procedural systems still contain not only procedural encodings, but straightforwardly non-procedural encodings as well. The distinction in practice is not between declarative encodings and procedural encodings, but between declarative — truth-valued — encodings, on the one hand, and representational encodings that do not carry truth values — features, categories, objects, and so on, on the other hand. The issue is not the dispensability of declarative encodings in favor of procedural encodings, but the dispensability of declarative encodings in favor of procedural encodings *plus* non-truth-valued encodings. Procedural encodings are in common to both sorts of system.

The debate in practice concerns the nature and dispensability of truth-valued encodings, but general encodingism is so deeply presupposed that the presence of non-truth-valued encodings in "strictly procedural" systems is not taken to be of relevance. Among other consequences, what are taken to be strictly procedural systems or models of procedural semantics are, for this reason as well as others, still encoding models, and, therefore, quite distant from an interactivism model (Miller & Johnson-Laird, 1976).

PROCEDURAL SEMANTICS

Procedural semantics (Hadley, 1989; Woods, 1986, 1987) is related to the proceduralism involved in the procedural-declarative controversy, but it shifts primary focus away from proceduralism as a purely *operative* consideration, that is, it shifts away from procedural meaning strictly in terms of what can be done *with* a symbol, to proceduralism as involved in *truth criteria* for symbols. In particular, in standard model theoretic semantics (including its variants of possible world semantics and situation semantics) meanings are construed in terms of maps to extensional sets. These maps may be from symbols to elements or sets in the actual world, or there may be distinct maps associated with each symbol that maps each possible world to a set in that possible world, or analogously for situations.

The critical characteristic that all such approaches to semantics share is that the crucial maps are construed in the pure extensional maps-as-sets-of-ordered-pairs sense, with no attention to how those correspondences are to be computed, detected, instantiated, or otherwise

realized in actual system processes — no attention to how the semantics is supposed to actually happen in the system. In this sense, *semantic competence* is taken to be unproblematic for the foundational issues of semantics. In the *Tractatus*, Wittgenstein made a similar assumption, and later recognized how untenable it was. Contemporary encodingism, with sparse exceptions, has still not understood this point.

Procedural semantics focuses on this point, and attempts to provide an account of how symbols are grounded in actual system processing with respect to the world. There are many interesting issues that arise in this attempt, such as those of the issues and design consequences of uncomputability, the dangers of empiricist verificationism, the characterization of abstract procedures, and so on.

We are in strong agreement with the process and procedural insights involved in this approach. Any model of representation must ultimately be made good in terms of real systems and real processes within those systems, and to ignore or dismiss such considerations is short-sighted and unwise in the extreme. Any model of representation that is impossible to instantiate in real systems is impossible as a model. Procedural semantics has hit a point of serious vulnerability in standard approaches.

Still Just Input Correspondences

Procedural semantics, nevertheless perpetuates the fundamental errors of encodingism. The criterial procedures compute characteristic functions of input categories. These characteristic functions may not be effective, and may only serve to defeasibly ground or constrain denotational assignments of a symbol to the world, rather than constitute in some full sense the meaning of that symbol. Such specifics are important for the procedural semantics project, but they do not alter the encoding assumptions that the basic level symbol meanings are constituted as correspondences with what they encode. The proceduralism is concerned with how any such encoding correspondences could be computed in real procedures, with all of the problems of uncomputability, and so on, not with any basic flaw in encodingism per se. Proceduralist characteristic functions are differentiators, and there is no basis for assuming that they represent what they differentiate. To assume otherwise is to presuppose encodingism.

Given all of procedural semantics, then, it still offers no answer to the question of how or in what sense the system knows what its

procedures have just detected — of what is on the other end of the correspondence just computed. It offers no solution to the problem of representational content. It offers no solution to the problem of error: a procedural detector will detect whatever it detects, and only an external observer could claim that that associated symbol is *supposed* to represent cows, even though what it just detected is in fact a horse on a dark night (see the disjunction problem discussion below).

Transducers. In Fodor's more recent conception of a transducer — as a generator of encodings that does *not* involve inference — such characteristic-function procedures are transducers, and this version of a *semantic* transducer is no more possible outside of an observer semantics than any other version of transducer. The notion of a transducer as a generator of encodings presupposes that the system state generated by the transducer somehow represents for the system what is on the other end of the correspondence relationship instantiated by the transducer process — and that, as should by now be clear, is impossible. It also presupposes that the encoding generated by the transducer represents the "right one" among the unbounded correspondences that any such instance of a correspondence will be associated with — the retinal processes, the light patterns and flux, the electron transitions in the surfaces of objects, the histories of the creation and movements of those objects, the objects or their properties themselves, and so on — and it presupposes that *errorful* such correspondences (how can they even exist?) can somehow be characterized as such for the system itself. These problems, of course, have not been solved, and cannot be if the encodingism critique is correct. The introduction of considerations of procedural computability does not alter any of these issues.

Still Just Inputs. One immediate indication of the gulf between procedural semantics and interactivism, in spite of their common recognition of the critical importance of process and functional ontologies, is that procedural semantics is still presumed to be definable in terms of the processing of *inputs.* There is no necessity of action or interaction in this view. The core of interactive representationality, of interactive representational content, is the indication of potential interaction. Such indications may be *based* on interactive differentiations of the environment (or may not be), which, in passive versions, will be similar to procedural semantics computations (but with timing and topological considerations added), but the factual correspondences with

the environment that are instantiated by such differentiations do not and can not constitute representational content.

Content for the System. Such factual correspondences (causal, transducer, tracking, procedural, connectionist, and so on) are relations *between* the system and the environment; they are not relations or states *in* the system. Content, on the other hand, *bears* relations to the environment — such as truth or falsity, or, more generally, "aboutness" — but content is itself strictly *in* the system. Otherwise, if content were not *in* the system, then content could not be either functionally or epistemically efficacious (or real) *for* the system. What a representation is *about* **in the world** is not in the system, and, therefore, cannot in general be efficacious internal to the system, yet content *must* be efficacious for the system in order to *be* content for the system. In general, then, relations between system and environment cannot be content.

Although content itself is not and cannot *be* a relation to the environment, content can and does *have* relation to the environment. Content has relation to the environment in the *implicit* sense that an indication of a potential interaction is implicitly a predication of an implicitly defined class of interactive properties to the environment — content is *in relation* to the environment in the sense of being *about* that environment. Procedural semantics, in contrast — like other encodingist approaches — assumes that some sort of proper correspondence relations to the environment will constitute or provide content about the other ends of those correspondences.

Procedural semantics, then, is pushing at the edge of contemporary approaches to semantics in ways that we think are insightful and valuable. But it cannot escape encodingism so long as representation is presupposed to be some form of correspondence, so long as representation is assumed to be generated or generable by the processing of inputs per se — and so long as the necessary involvement of action and interaction, and the modal involvement of potential interaction, is not recognized.

SITUATED AUTOMATA THEORY

Rosenschein and Kaelbling (Rosenschein, 1985; Rosenschein & Kaelbling, 1986; Kaelbling, 1986) propose an alternative to the usual approach to "knowledge" in machines. In standard practice, machines are designed in terms of internal operations on internal symbols that can be

interpreted as being about the world — the familiar symbol manipulation approach. Instead, they propose a "correlational" approach in which machines are designed such that the internal states have provable correlations with, or correspondences to, environmental conditions. There are no manipulations of internal data in this approach, and an internal state that corresponds to a complicated environmental condition will not in general have any internal state structure that corresponds to those complications.

Note first that being in an internal state is the only internal functional reality that a machine can have. Even in the case of manipulated structured internal data, in a strictly functional sense, that data does no more than participate in the constitution of a single overall functional state — and its differentiation from other states — from which the machine can proceed to still other states. The automata perspective is always applicable to any discrete machine. Rosenschein and Kaelbling's situated automata approach, then, is a move away from all of the internally structured data complexities of the usual symbol manipulation designs to a purely functional approach in which that functionally superfluous state structure is eschewed in favor of more direct correlations between unstructured states and environmental conditions.

The formal approach that they offer has something of the flavor of classical recognizer automata theory (Brainerd & Landweber, 1974; Hopcroft & Ullman, 1979; Minsky, 1967). The focus of consideration, however, is not on the recognizable input symbol strings, but on the transitions among differentiable classes of environmental conditions that correlate with transitions among the automata states. That is, there is a move from the input strings per se to the environmental, the situational, conditions associated with those input strings. Thus the term "situated automata." They have developed both a formal logic for *defining* machines in these terms and a programming language for *designing* such machines in terms of such definitions.

The functionally superfluous internal structure of standard approaches, of course, is generally interpreted as corresponding to the propositional, or truth functional, structuring of the environmental conditions being represented by that data. The differentiation of that structure may be superfluous to the functioning of the machine, then, but it would usually not be considered superfluous to the semantics of the machine's representations. The functional equivalence and greater simplicity of the situated automata approach is one more demonstration of

the sense in which such semantics, with all of its structuring and complexity, is of and for the designer, not for the machine.

Moreover, Rosenschein (1985) points out that in the symbol manipulation approach, the knowledge state of a machine depends on the interpretational attitudes of the designer or user, leaving open the possibility of the same machine in the same state being interpreted as having different knowledge states by different users. Knowledge state in this approach is not an objective property of the machine. The correlations between machine states and environmental states of the situated automata approach, however, *are* objective properties of those environmentally situated machines.

Rosenschein (1985) is quite clear about the fact that the correlational "knowledge" in a situated automata's states is purely epiphenomenal — it is strictly for and from the perspective of the designer (though it is not designer-*relative* as is the case for interpreted data structures). He is equally clear that the representations of standard data structures are designer representations, and are not so for the machine. The use of such terms as "knowledge," however, and "awareness" (Kaelbling, 1986), tends to obscure that point, which recedes in explicitness from the 1985 paper onwards.

"Epiphenomenal" or "implicit" are exactly the correct terms for the "knowledge" in these situated automata state correlations with their situations. Situated automata may be *active* in the sense of emitting outputs, and even *re*active in the sense of responding to unpredicted changes in inputs, but they are not *inter*active in the sense of generating or producing outputs *for the sake of* (from the machine perspective) their subsequent input or environmental consequences. Situated automata theory replaces interpreted truth-conditional correspondences with causal correlational correspondences, but in neither case is there any true interaction or goal-directedness, and, therefore, in neither case is there any emergence of representational content — content that is, *for-the-machine*, right or wrong.

Rosenschein and Kaelbling's primary concern is machine design, and, from that perspective, our points about what their approach does *not* do take on less import — many design goals can be accomplished without taking the interactive considerations explicitly into account. By embedding a situated automaton in the perspective of interactivism, however, we can see that a situated automaton is essentially a passive "interactive" differentiator, possibly with *outputs* — not procedures or

strategies — contingent on the differentiation categories. Correspondingly, their design language and logic are of such passive differentiators: they differentiate desired environmental conditions — conditions in which particular outputs are intended to be emitted in the case of a robot (Kaelbling, 1986, 1992). The restrictions of situated automata theory, therefore, not only do not account for the emergence of representation (which it was not intended to address in the first place), but, even from a design perspective, the passivity and lack of goal-directedness limit the design power of the approach (see, however, Kaelbling, 1992). Reactivity is intrinsically of less power than goal-directed interactivity, both for "detection" of environmental conditions, and for changing those conditions.

NON-COGNITIVE FUNCTIONAL ANALYSIS

Hatfield and Kosslyn (Hatfield, 1986, 1987; Kosslyn & Hatfield, 1984) argue for the existence and the usefulness of a level of analysis that is above the strictly implementational level — neurophysiological, in the case of human beings — and below the level of symbol manipulation computations. They call this non-cognitive functional analysis.

Although the case for such a level is made primarily in terms of reviews of numerous theories, mostly in the psychology of vision, the *general* point that such a level exists is made by the existence of automata theory. This is precisely a functional level of analysis, abstracted away from implementation, but not involving symbol manipulation. It is not the only such form of analysis, but it is a form that is always, in principle, applicable to any finite discrete machine.

The issue, of course, is whether such a level of analysis can be useful. Automata theory is generally avoided in favor of Turing machine theory, or some easier to work with equivalent, such as a programming language, because of the limitations of computational power of automata. On the other hand, a Turing machine is itself nothing more than a finite automaton with an unboundedly extendable memory. The real appeal here seems to be that the elements in that memory — on the Turing machine tape — are generally given interpretations as symbolic encodings. There is nothing in Turing machine theory per se that assumes or requires this interpretation of its tape (though such an interpretation was involved in Turing's motivating model for the theory). This makes it clear that the underlying paradigm is not Turing machine theory per se, but the standard computer metaphor with its symbol manipulations. All

of the issues of epistemology and semantics are thereby smuggled in with a slight of hand — under the guise of favoring Turing machine theory over automata theory because of its higher *"computational"* power.

The task of demonstrating the usefulness of a noncognitive functional analysis, then, faces formidable inertia and opposition. Here too there is help from examples already existing — this time in various models of vision, and (for perhaps the clearest demonstration of existence) in Artificial Intelligence in the form of connectionism. It may well be possible to assign representational content to the outputs, and perhaps the inputs, to a connectionist network, but there is in general no coherent way to assign content to the activations and influences *within* the network. Nevertheless, the function of the network can itself be analyzed, just not in symbol manipulation form.

The connectionist example raises the possibility of a kind of external observer semantics, in which the observer or user assigns representational content to the inputs and to the outputs, with no such assignments internal to the system processes. Some machines may be most useful to a user in terms of such an external observer semantics; certainly this is the pragmatic stance of most users to most computers. More fundamentally, it is not logically *necessary* for useful observer-representational machines to have their internal processes decomposed in such a way that some of them can be designated as processes operating on others as symbols. Nevertheless, it is clear that a user semantics — internal *or* external — simply avoids the fundamental issue of representation.

Instead, noncognitive functional analysis suggests an explication of representational content in terms of the *functioning* of the overall system. The general notion seems to be that "**X** represents **P**" if **X** in the system serves to influence the system's processing so that the system operates in accordance with **P** existing, or being the case. Determining what counts as functioning in accordance with **P** is not as clear as would be desired, but there seems to be a reliance on notions of goal-directedness and biological adaptedness here — functioning in accordance with **P** is functioning in such a way that goals are approached and adaptedness is maintained by processes that rely on **P** for that process's functional appropriateness. "Functional" here is being used in the sense of "serving a function," not necessarily in the sense of "computing a function."

This notion of representation has intuitive appeal, and not only to Kosslyn and Hatfield (for example, see Bogdan, 1988a, 1988b, 1989). It would certainly seem to be a requirement of representing something for a system that the system thereby comes to be able to function (better) in accordance with that something, as obstacle or resource or whatever.

The Observer Perspective Again

But these considerations are all from an observer's perspective. They leave untouched the issues of the emergence of representation and representational content for the system itself. To illustrate, the purely functional notion above, if taken as *sufficient* to representational content, would yield the conclusion that a thermostat has full representations of temperature — the functional considerations and the adjustment of functioning in accordance with environmental conditions are both present. Similarly, in Hatfield's examples, various activities in the visual nervous system are said to be representing properties of light — just not in a symbolic form. Again, in the strictly *functional* sense, this is a misnomer — but not otherwise problematic — but it is *not* an explication of the emergence of representation *for the system*.

This point does not necessarily count against Hatfield and Kosslyn's general arguments, because they are arguing for a form and level of analysis for the *analysis of the functioning* of psychological systems, and issues of representational emergence for the system itself may not be relevant to the concerns of particular such analyses. It is not clear, for example, that *anything* in the activities of the optic tract per se serve as representations *for the organism itself.* Nevertheless, the functional roles of some of those activities might well be analyzable in the sense of noncognitive functional analysis. This is, in effect, a kind of evolutionary design perspective analysis, asking "Why is this here?" and "Why does it do what it does?" In spite of this, it is regrettably confusing that such analyses be discussed in terms of "representation" — the only representations are for the psychologist, not for the system.

Underlying this form of analysis is the notion of various activities and states of the system having, or attaining, factual correspondences with environmental conditions, and influencing the further activity of the system in such a way that that activity is functionally "appropriate" to those conditions. A formal example of this conceptual approach is situated automata theory.

Such factual correspondences between something within the system and something in the environment can occur and come into factual existence — and can appropriately influence further activity — without there being any *flow or transmission* of such "correspondence." In situated automata theory, a state in an automaton can be entered reliably only when certain conditions in the environment obtain without there being any "flow" of such correspondences into that state. An interaction, for another example, might differentiate an environmental condition by the *overall pattern* of the interaction without any *part* of the interaction constituting such a differentiation. Flow or transmission of representation or information, therefore, is not required — in any sense of the terms. This is unlike the transmission and progressive processing of "symbols" in a standard information processing model.

This point seems to be missed in at least one part of Hatfield's discussion (1987), perhaps because of a confusion between functional analysis in the sense of *serving* a function and functional analysis in the sense of *computing* a function, though most likely simply because it is such a dominant manner of thinking in the psychology of vision. He suggests:

> Such a psychology should be acceptable to direct theorists [Gibsonians], in that it avoids cognitivism. Its acceptance by direct theorists would allow them to discuss the flow of information within information pick-up devices using a functional, rather than a neurophysiological, vocabulary. The notion of *representation* would allow them to chart the flow of information beyond the retina, and the notion of *computation* would allow them to give an account of how higher-order stimulus information is detected. (1987, pp. 41-42)

This seems to be wrong both in its presumption of the *necessity* of accounting for such "flow," and as a misunderstanding of Gibson in the notion that any such account would be *compatible* with Gibson. Information does not, and need not, *flow* at all in information pick-up (Bickhard & Richie, 1983).

Aside from its relevance for understanding Gibson, and vision, this point is also one more illustration of how difficult it is to *not* treat representation as correspondence, and correspondence as encodings. Correspondence, causal or informational, can help *explain* how a representational system works and how it is successful, *from the perspective of an external observer* (a psychologist, perhaps) *on both the*

system and its environment, but it is the wrong category within which to differentiate the nature of representation per se. Noncognitive functional analysis contributes to this task of functional explanation, but participates in the encodingist confusion about representation itself.

There is an interesting almost-convergence between noncognitive functional analysis and interactivism. Interactivism models the emergence of representation *as* a function *within* functional interactive systems. The general perspective of functional analysis, therefore, is shared. To leave the analysis only at the general functional level, however, without an explication of the *representational* function per se, yields at best a pragmatics of the overall system, with no representation and no semantics. It leaves all actual representation solely in the person doing the analyses, and, thus, leaves the nature of such a being who can represent and do such analyses still utterly mysterious.

BRIAN SMITH

Brian Smith (1985, 1987, 1988), in addition to his critique of Lenat's project (1991), has tackled some of the foundational problems inherent in contemporary approaches to knowledge and representation. He explores and attempts to sort out the amazing tangle of confusions, ambiguities and equivocations among approaches involving linguistic expressions, model theory, implementations, interpretations, programs, processes, specifications, the use-mention distinction, and transitive and intransitive correspondence relations. All of these seem to capture at least some aspect of representation or knowledge, but identification of the aspect with the whole and confusion between cases seems rampant. Smith is to be commended, though perhaps not envied, for attempting this Sisyphean task.

Smith distinguishes between the *functional role* of representations — the sense in which they must have some sort of functional role for the epistemic agent involved — and the *representational import* — the sense in which they must somehow be about the world. Furthermore, he recognizes that the standard explication of representational import in terms of correspondence cannot be correct. Among other considerations, representational import must be capable of being *wrong* — unlike, for example, the "mere" correspondence between rising sap in maple trees and the weather (1987, p. 4).

To this point, we are in strong agreement with Smith, but, from the interactive perspective, his analysis, nevertheless, seems still caught in

the maze that he attempts to transcend. The point that representational import must be capable of being wrong is a deep and correct insight, but it does not explicate far enough. If it is simply left as the capability of being wrong *per se*, then interpretation is open to the possibility of "representation" being wrong *to the observer*, for example, and we have a full observer encodingism. Representational import must be capable of being wrong *to the epistemic system itself*. It is this requirement that is not captured in Smith's analysis.

Correspondence

Smith characterizes representation as narrower than correspondence, but, nevertheless, as a species of correspondence, a differentiation within the general phenomena or domain of correspondence. He claims that "correspondence is ... more general ... than representation" (1987, p. 30). Representationally relevant correspondences are, ultimately, correspondences with the (rich) ontology of the world, which thereby provides the ground for representation. But this is a standard definition of an encoding in terms of what it represents, in terms of what it has a known correspondence with. It fails to address how a correspondence could have such epistemic properties, how such a correspondence could be known, how it could be known to be right or wrong, how it could be specified what such a purported correspondence was supposed to be with, and so on — the full array of encodingism circularities and incoherencies. Correspondence is simply the wrong framework for understanding the *emergence* of representation, however much it might in some cases be involved in the *explanation* of representation. Identifying correspondences may be useful for the observer, but it is not the basis for representation in the epistemic agent.

Smith contends that the functional role of representations and the representational import of representations must somehow be *integrated* into what he calls the *full significance* (1987, p. 5) — they are *not* independent. Interactivism suggests that he is quite correct in this, but that the *presuppositions* involved in Smith's notions of functional role and representational import block this integration.

Participation

Smith (1988) generalizes the notion of Turing machine conceptions of computation to a more general notion of systems that are *participatory* in their environments, such as in the case of clocks being

participatory in the flow of time. He points out that clocks are representational, at least for users, without engaging in the symbol manipulations of standard conceptions, and that computation in the usual sense of manipulations on symbols can only be defined in terms of prior notions of semantics, rather than providing the arena within which semantics itself can be modeled (for a convergent point, see Bickhard & Richie, 1983). We are in full agreement with these cracks in standard Turing machine inspired frameworks, but would point out that Smith still ends up, even in the case of clocks, with representation via correspondence. In this case, it is temporally extended, participatory, correspondence with the flow of time itself, but it is nevertheless still an encoding notion of representation. Smith acknowledges that clocks represent only for their users, and that the ultimate goal is natural representation, as it occurs in humans. We claim that aspiration is impossible to fulfill unless the encoding framework itself is transcended.

No Interaction

In particular, there is no intrinsic notion of *inter*action involved in Smith's notion of functional role, nor even of *environmental* action at all — functional role in terms of further system activities strictly internal to the system itself, such as drawing inferences, seems to satisfy Smith's notion here — and, therefore, certainly no intrinsic involvement of *goal-directed* interactions. But, if the interactive analysis is correct, then representational content — representational import — that is capable of being wrong for the epistemic system itself emerges only in goal-directed interactive systems.

Similarly, there is nothing akin to open ended interactive differentiation in Smith's notion of correspondence. Representation is not just correspondence with the addition of the epistemic-ness of that correspondence — that's just encodingism — but by attempting to finesse the question of what supposedly makes some correspondences representational and some not, Smith is ultimately committed to such a position (e.g., 1987, p. 34).

In the interactive view, representation does not emerge in knowledge of what the differentiations are differentiations of — are in correspondence with — but instead representation is emergent in predications of the potentiality for further interactive properties. Such predications of interactive potentiality will often be *evoked* by — be contingent upon — instances of environmental differentiations, such as a

frog predication of "eating opportunity" evoked by the factual differentiation of a fly. In such an instance, however, the representational content — the potentially system falsifiable content — is of "eating opportunity," not of "fly." The factual correspondence with the fly serves a functional role in evoking the representation, the predication, of "eating opportunity." The factual correspondence does not constitute that representation.

Furthermore, even though in some paradigmatic cases the *contingencies* for such predications will involve factual, perhaps causal, correspondences — such as from retinal image to light pattern to certain properties of the physical world (though even this simple correspondence model ultimately does not work, Bickhard & Richie, 1983) — in general, representational contingencies need not involve such correspondences. An interaction, for example, may *create* the further interactive potentialities that its outcome indicates to the system, not just *register* or *detect* those potentialities. The interactions of opening a can of cola, for example, or filling a glass with water, *create* the potentialities for taking a drink; they do not merely detect them. Furthermore, making the distinction between such detections and creations *from the perspective of the system itself* is very difficult — it constitutes the epistemic agent's functional transcendence of solipsism, and it involves the differentiation of self from world that occupies the first several years of infant and child development. Even in cases where (at least to an adult human being) there is a clear case of the *creation* of interactive conditions, and, thus, no correspondence with what is represented involved in the initial creating interaction, there is, nevertheless, still representational content in the ascription of the created interactive properties to the world. Interactive representation may factually involve correspondences, and may in some cases be explained in terms of such correspondences, but those are not known correspondences, and such correspondences are not necessary.

Correspondence is the Wrong Category

Fundamentally, correspondence is the wrong approach to representation because correspondence serves to pick out, to specify, what is to be represented, and to define representation as having the function of representing what is thereby specified. This conflates the functions of epistemic contact with the world — differentiation — and knowledge, representation, about what that contact is with — representational content. Encodingism *defines* epistemic contact *in terms of* such knowledge of

what the contact is with — it does not differentiate the two. Encodingism *requires* such knowledge of what contact is *with* in order to have such contact at all. This is not problematic for real stand-in encodings, since the knowledge and the contact are provided simultaneously in the provided, the defining, representational content. A correspondence approach to representation, however, and a strict encodingism, do not allow their differentiation; both are required before either can be obtained — this is the encodingism circularity and incoherence. Interactivism *separates* epistemic contact from knowledge about what the contact is with, and does not require both before either is possible.

Still further, interactive representational content is not representation ***of*** what has been differentiated, but only representation ***about*** or that ***follows from*** what has been differentiated. That is, it is representation of various further interactive properties indicated by the relevant differentiation. It is always *defeasible*, and it is always *partial*. The assumption that those further interactive properties fully specify or individuate what the differentiation is a differentiation *of* is an *additional* claim that is *not an intrinsic aspect of the representation itself* — even though such a claim might, conceivably, in some cases be correct. In general, however, there is always more to learn about what differentiations are differentiations *of*, and there are always further relevant subdifferentiations that can or could be made.

We agree with Smith in his concerns about the confusion in the contemporary literature, with his contention that representation involves an integration of import and function, and with his insights about representational import involving more than correspondence — in particular, that it must be capable of being wrong. We would add "It must be capable of being wrong *to the system*." (Bickhard, 1993a, in preparation-c) and contend that Smith's restriction to representation as a form of correspondence — and his related neglect of interaction and goal-directedness — makes the emergence of such "wrongness," and, therefore, the integration of import and function, impossible.

ADRIAN CUSSINS

Cussins has proposed an important version of the involvement of "appropriate functioning" in representation (1990, 1992; Clark, 1993). He proposes "non-conceptual content" as the most primitive form of representation, a form upon which more standard "conceptual content" is developed. Non-conceptual *content* is content specified in terms of non-

conceptual *properties*; non-conceptual *properties*, in turn, are those that can apply to an organism without that organism necessarily having concepts with that content itself. "Orienting north" can apply to a paramecium without that paramecium having any concepts at all, and certainly not a concept of "orienting north." A conceptual property, in contrast, can not apply to an organism without that organism itself possessing the concepts involved. The property of "thinking of someone as a bachelor," for example, could not apply to someone unless that person had the concepts appropriate to the property bachelor, such as "male," "adult," and "unmarried." "Thinking of someone as a bachelor," therefore, is a conceptual property.

The critical notion here for our purposes is that of non-conceptual content. The shift from non-conceptual representation to conceptual representation, of course, is of fundamental interest and importance, but the most important move is the attempted naturalization of content — of any kind — via non-conceptual content. As the "orienting north" example illustrates, some contents can consist in certain action dispositions and abilities. Being able to successfully negotiate particular domains can instantiate non-conceptual properties.

Representation as instanced in such action and interaction capabilities is, clearly, a move at least partially convergent with the interactive model of representation. Such interaction capabilities involve indications of what will work for the organism under what conditions local to the organism. Such indications, in turn, will have truth conditions — conditions under which they are correct, and conditions under which they are not — even if those truth conditions are not explicitly represented.

The space of possible such indications will constitute a kind of local frame for action and interaction. They will necessarily be organism centered, and action focused: they will be indexical and deictic. The "necessity" here is simply that the particular organism involved is the only privileged origin that is intrinsically available to that organism: any other frame for possible action — for example, a Cartesian spatial frame within which the organism has some location — requires some sort of perspective from outside the organism, a perspective that is not available to the organism. (One exception, perhaps, might be the limit of constructions of more and more context invariant frames, beginning with system-centered frames — something that can sometimes be quite useful to do for organisms that are capable of it, such as human beings).

Furthermore, the environmental conditions upon which interaction indications are based will necessarily be implicitly represented — to assume explicit representations of them, at least at a foundational level, is to assume an encodingism — and the truth conditions involved in the interaction indications will similarly be implicitly represented. A consistently developed recognition that representation is intrinsically emergent in interaction systems — in pragmatics — forces such indexical, deictic, implicitness in order to avoid the failures of encodingism. Cussins' model constitutes a move from input-correspondences to input-correspondences-plus-appropriate-functioning to successful-interactive-functioning-per-se as the locus or domain in which representation is emergent. In moving to a pragmatic locus for representational emergence, Cussins has diverged significantly from standard approaches, and has converged in important ways with the interactive model.

Cussins, however, does not develop the interactive notions of implicitness, of having truth conditions without representing them, or of indexicality and deicticness. Most importantly, he does not develop the necessity for indications of potential interactions to involve associated indications of possible internal outcomes of those interactions. Without such indications of internal outcomes, there is no way for the interactions to fail from the perspective of the organism itself, and, therefore, no way for the implicitly defined truth conditions to be falsified from the perspective of the organism itself — and, therefore, no way for there to *be* any implicitly defined truth conditions for the organism itself. Without such a possibility of error detectable by the organism itself, there can be no genuine representation for that organism (Bickhard, in preparation-c).

Cussins, then, has moved a long way toward a pragmatic, interactive, model of representation. We suggest, however, that several additional characteristics of interactive representation must be explicitly modeled before that move can be complete. Most importantly, indications of interaction outcomes, and the implicit conditions and truth conditions that are involved in such indications (Bickhard, 1992c), are required for there to be a model of representational content that is naturally emergent in the organism and for the organism.

INTERNAL TROUBLES

In recent years, a number of difficulties inherent in encodingist approaches to the nature of representation have become focal issues in the

literature. There have been many attempts, and much discussion of those attempts, to understand those difficulties, and to solve or avoid them. In these attempts, the difficulties are taken as issues about representation per se, issues that must be solved in any ultimately satisfactory model of representation. They are *not* taken as reductios of the basic encodingist framework for approaching representation. For our purposes, however, these difficulties illustrate even further the morass of impossibilities that encodingism leads to, since, we argue, *none* of these difficulties *should* be difficulties at all — the issues either do not appear at all, or do not appear *as difficulties*, in the interactivist perspective. Conversely, they *cannot* be solved from within encodingism. These problems are purely internal to encodingism, and focusing on these problems serves simply to distract attention from the underlying encodingist presuppositions that give rise to them.

Too Many Correspondences

The notion of representation being correspondence-plus-functionality has appeal — especially if functionality is construed in terms of the system's goals or in terms of the evolutionary history of the species. Functioning, after all, is what we do with representation, and appropriate functioning would seem to be the obvious candidate for picking out what a representation is a representation of.

This is especially powerful if we consider that any correspondence between states internal to the system and conditions external to the system is also going to participate in unbounded numbers of additional correspondences: with light patterns, with electron interactions in the surfaces of objects in the visual field, with aspects of past histories of those objects, and so on. One problem with encodings as correspondences is simply that there are too many correspondences — any particular correspondence drags along unbounded numbers of associated correspondences (Coffa, 1991). Appropriate functioning might be one way to select which is the relevant correspondence.

There is a sense in which this is correct: an observer analyzing a system and noticing multiple correspondences of internal states of the system might select among those correspondences for the one with respect to which the system's behavior was "appropriate," and, on that basis, conclude that that internal state encoded whatever was on the other end of the selected correspondence — some particular object, say. The observer might even be "correct" in some evolutionary sense. But the

analysis takes place entirely from within the observer perspective, not the system's perspective. The "selection" that takes place is an act that the observer engages in — selecting among all of the correspondences that the observer can find, the one that seems to be most "appropriate" to the actual system functioning. At best, the analysis picks out things in the environment that are *functionally* relevant for the system; things with respect to which the system functions in that environment.

In order for such functioning to pick out, to select, appropriate representational content *for the system*, the system would have to already have representational content for *all* of the correspondences, among which the functioning could (somehow) then select the "right" correspondence. In other words, the only way that correspondence plus functioning will get the right representational content is for the representational content to be already present. This prior presence of representational contents is presupposed for an actual observer — after all, the observer can "see" all those elements in the environment, and can track or analyze the causal (for example) chains among them that give rise to the correspondences. But the system would have to be in an observer position with respect both to itself *and* with respect to all of the corresponded-to things in its environment in order to be able to engage in a similar "selection" of appropriate representational content. As a model of representational content, this is merely the by now familiar circularity of encodingism: the content has to already be there in order to account for content.

This approach, then, does not even address the issue of the constitution or emergence of representation for the system itself. It leaves a mystery how the system could represent, could know in any sense, what *any* of its correspondences are with, or that there *are* any such relationships as correspondences. Correspondence is not representation, and adding functioning does not make it so.

Disjunctions

Another problem from within the encodingist perspective is called the disjunction problem (Fodor, 1987, 1990; Loewer & Rey, 1991; Bickhard, 1993a). It is a version of the general problem for encodingism of how error can be defined for the system itself. The disjunction problem follows from the fact that, if representation is taken to be constituted as correspondence, even correspondence plus functioning, then what we would want to call *errors* of representation will have

exactly the same properties of correspondence with system states as will *correct* representations. A common example is to consider a representation for "cow" which is correctly evoked by instances of a cow. On a dark night, however, it might also be evoked by a horse. We would like to classify this as an error, but it is not clear what could block the alternative conclusion that our original representation is simply a representation of "cow or horse (on a dark night)" instead of a representation of just "cow." What makes some evocations — some correspondences — correct and others in error?

Fodor has proposed one solution which he calls the "asymmetric dependence condition." The basic notion is that what we want to count as errors should be in some sense parasitic on the correct correspondence evocations. He attempts to capture that parasiticness with the claim that the possibility of errorful evocations is dependent on the possibility of correct evocations, while the reverse is not so: errorful evocations by horses are dependent on the possibility of evocations by cows, but evocations by cows are not dependent on the possibility of evocations by horses. Thus, there is a dependency between correct and incorrect evocations, but it is asymmetric. This asymmetry, therefore, is proposed as differentiating what is supposed to be represented from evoked correspondences that are in error.

Aside from technical problems with this proposal (Bickhard, 1993a; Loewer & Rey, 1991), we point out that it is an analysis, again, strictly from within an observer perspective. At best it would differentiate representations from errors for an observer. It provides no way whatsoever for a system to make such distinctions for itself, and, therefore, no way for a system to distinguish error for itself. A system would have to already know what its correspondences were with independently of the encoding correspondences at issue in order for such modal asymmetric dependencies to tell it what its representations were representations of — assuming contra fact that the system could analyze any such modal asymmetric dependencies in the first place. Once again, there is no approach to the problem of genuine representations for the system itself, not just in the view of an external observer.

The observer dependency of this notion is illustrated by a counterexample: consider a transmitter molecule that, when it docks on a receptor in a cell surface, triggers various functional activities inside the cell. Now consider a poison molecule that mimics the transmitter molecule, thereby inappropriately triggering those internal-to-the-cell

functional activities. There is an asymmetric dependence between the ability of the transmitter molecule to initiate the cell activities and the ability of the poison molecule to do so, yet neither the transmitter nor the poison nor the internal cell activities are representations or are represented. In fact, there is nothing epistemic going on here at all — cells, in general, are not epistemic agents. The proximate activities in the cell that are triggered by the docking into a receptor molecule are functional for the cell in corresponding to the external transmitter molecule, and in thereby corresponding to whatever initiates the release of that transmitter molecule elsewhere in the organism. The asymmetric dependence exists here, but it is an asymmetric dependence at a strictly functional level, not at an epistemic level. Only from an observer perspective can those internal activities be construed as encodings of the transmitter or its normal conditions of release. Fodor's asymmetric dependence criterion can at best capture an observer dependent functional distinction. It does not suffice for any epistemic relationship at all — at least not for the system itself.

It should be noted that this general example of a transmitter molecule triggering corresponding functional activities inside the cell is also seriously problematic for general "correspondence plus function" approaches to encodingism. This point holds whether the functioning is taken to be general computational functioning (e.g., Smith, 1985, 1987, 1988 — see above) or teleological functioning (e.g., Dretske, 1988) (or teleological functioning *independent* of such correspondence [Millikan, 1984]): the cell activities instantiate both. This is an example of 1) correspondence, tracking, from inside the system to outside the system, with 2) appropriate further normal activity that depends on that correspondence, and 3) which correspondent-dependent activity is itself adaptive and the product of evolution (and has as its evolutionary function to track and respond to such transmitters). Yet there is no representation involved at all — at least not in any epistemic sense for the system — only a mildly complicated set of functional relationships.

Wide and Narrow

Still another difficulty that encodingism has encountered recently is an intrinsic context dependence of what an encoding correspondence is with, and, therefore, of what an encoding represents. The problem with this discovery of intrinsic context dependence is that *encodings* are defined in terms of their representational content, in terms of what they

represent, and, therefore, in terms of what they are in representational correspondences with. If those representational correspondences are themselves indeterminate, and instead relative to context, then it is not clear how to model encodings at all.

One source of recognition of the difficulty is the Twin Earth problem. Imagine a twin of earth, a twin down to every particular, including the individual human beings, except that what we call "water" on this earth, H_2O, is instead XYZ on twin earth, where XYZ is a different chemical, but is otherwise indistinguishable with respect to flowing, being drinkable, supporting life, and so on. The point is that even though the conditions inside each person's head are identical to those inside their twin's head on twin earth, nevertheless what is represented by "water" is different on the two planets. There is an inherent context dependency.

Any moves to try to characterize representation in terms of lower level features, for example, say water in terms of flowing, being drinkable, supporting life, and so on, are simply subject to their own counterexamples in which they too are context dependent. If the representational content of representations is supposed to be something that those representations are in correspondence with, then such context variability robs encodings of any determinate content, and, therefore, makes them not encodings — encodings are defined in terms of what they represent, and, as the twin earth example shows, what they represent is indeterminate.

This problem becomes even more pressing when context dependencies less extreme than those between earth and twin earth are recognized. Pronouns, indexicals, demonstratives, and so on are highly context dependent, and even names, even proper names, depend upon their context for determination of what they are taken to refer to. This context variability has generally been ignored or set aside as a special case, but the twin earth thought experiment shows that it cannot be ignored at any level.

One proposal for attempting to deal with this problem picks up on a proposal for dealing with demonstratives and other highly context dependent forms: construe them as invoking a *function* from context to encoded content, thereby capturing the context dependency (Bickhard & Campbell, 1992; Kaplan, 1979a, 1979b, 1989; Richard, 1983). Fodor (1987, 1990), among others, has adopted this strategy. The determination of the function in the head is called the *narrow content* of the

representation, while what that function picks out in a particular context is called the *wide content* .

Issues concerning narrow and wide content can get quite technical (Loewer & Rey, 1991), but the basic problem from the interactive perspective is relatively simple: Kaplan's original proposal was for certain language forms, and its plausibility depended on the presupposition that there were other context *independent* encodings that the context *dependent* functions could map into. For standard encoding views of language, that is a plausible presupposition.

For representation in general, however, it requires some ground of context independent encodings in terms of which all other narrow contents, all context dependent functions into content, can be defined. But if the issue infects *all* representation, then there can be no such ground of context independent encodings — no encodings defined in terms of context independently specified representational contents. This leads to an unspecifiability in principle of narrow content, and an air of mystery about what it could possibly be (Loewer & Rey, 1991).[8]

Red Herrings

From the perspective of interactivism, these problems are all red herrings. They exist as problems only for encodingism, and are manifestations of the incoherence of encodingism. They simply dissolve in the interactive model.

No one proposes that correspondences or correlations or covariations per se constitute representations. Yet the faith persists that some *special sort* of correspondence or correlation or covariation will be representational. The strategies for attempting to make good on this faith are all versions of attempting to add *additional constraints* on the class of covariational correspondences in order to narrow it down to the genuinely representational correspondences. These additional constraints range from "appropriate internal functioning" to "asymmetric dependence," from "causal transduction" to "generated by evolutionary selection." Even if one of such strategies *did* succeed in narrowing the class of correspondences to representational correspondences, this would at best be an *extensional* circumscription of representational correspondences, and would leave the *nature* of such representations, and representation in general, still untouched.

---

[8] Fodor (1994) attempts to do without narrow content. He does not address the more basic issues we have raised.

It is, in fact, quite easy to extensionally pick out the class of representational correspondences — they are *the genuine encodings*, such as Morse code or blue-print conventions or computer data codes. But these are all *derivative* forms of representation, and, as noted, do not touch upon the basic nature of representation. They all require an epistemic agent as an interpreter. Such genuine encodings, however — such genuine representational correspondences — do keep alive the Quixotic quest for encoding models of representation. They keep the red herring market in business.

Interactivism, in contrast, simply never encounters, is intrinsically not faced with, the problematics that force the exploration of so many hoped for solutions to the impossibilities of encodingism — the exploration of so many dead ends and blind alleys. The interactive model of representation does not enter into the circularities of presupposing representational content in order to account for representational content, and, therefore, is not forced into such epicycles in attempts to transcend those circularities.

For example, the context dependencies that seem to force a notion like *narrow content* (setting aside issues concerning the assumption that both language and cognition are encoded representations), are captured naturally and necessarily by interactive differentiators. What is in *fact* differentiated can in principle involve massive context dependencies, including the possibility of context dependencies on context variations that the system has never actually encountered, such as twin earth for humans, or BBs and pencil points for frogs, which cannot distinguish them from flies. Language, with utterances as operators on current social realities involves still additional levels of context dependency (Bickhard & Campbell, 1992). *Accounting* for such context dependency is trivial for interactivism because such dependency is inherent in the nature of interactive representation.

At the same time, the *existence* of such intrinsic context dependencies within the interactive model is not problematic for interactivism. Interactive representational contents are not defined in terms of what they are in correspondence with, but, rather, in terms of the indications of further potential interactions that are constitutive of those representational contents. Whether or not the implicit predications of those interactive potentialities are *true* will be potentially context dependent, but the implicit predication of those interactive potentialities will not itself be context dependent. The problems that have yielded

broad and narrow content distinctions are not problems for interactivism (Bickhard, 1993a).

Similarly, interactivism does not need "subsequent internal functionality" to pick out the correspondences that are representational for a system, because it is *indications of potential* functionality, of potential interactions — not correspondences — that constitute interactive representational content. Again, the problem is not so much solved by interactivism as it is that it simply never emerges in the first place.

The disjunction error problem arises solely because of the encoding identifications of representation with factual correspondences. If there are factual correspondences, how could they ever be wrong? Interactive representations are of potentials for further interactions, and the correctness, or lack thereof, of any such indication in a system is absolutely contingent — there is no problematicness concerning the possibility that such indications of potential interactions might be wrong. Interactive representation is not constituted out of factual correspondences with the environment, but out of contingent indications about the future. Furthermore, such indications are in the system, by the system, and for the system, and any errors encountered with respect to such indications are similarly of the system, by the system, and for the system. There is no uncashed observer perspective in this model (Bickhard, 1993a).

In general, the too many correspondences problem, wide and narrow content, and the disjunction problem are no more than symptoms of encodingism. They do not need to be solved. They *cannot* be solved within the encoding framework that produces them. They simply disappear, or, better, never appear, in the interactive perspective.

10

Representation: Issues about Encodingism

SOME EXPLORATIONS OF THE LITERATURE

In the literature of Artificial Intelligence and Cognitive Science, as in that of the relevant philosophical literature, there are sometimes recognitions — of various sorts and of various degrees of completeness — concerning the flaws and consequences of standard encoding notions of representation. Also as in the case of the philosophical literature, the proposed remedies and alternatives invariably reveal a remnant encodingism that vitiates the ultimate viability of the proposal. In this subsection, we will examine a sampling of these insights and proposals.

Stevan Harnad

Some Proposals. Harnad has developed a set of positions that have some parallels with, and a number of crucial divergences from, the interactive model and its associated critiques. He is in agreement with the interactive position that there are serious problems with standard notions of representation, and makes a number of proposals concerning them. Those proposals concern both the diagnosis of the problems, and steps toward their solution.

Avoid the Chinese Room. A primary entree into Harnad's positions is through consideration of the classical Turing test, and of Searle's Chinese room argument against it. Harnad takes it as a primary negative task-criterion to find a model that is not vulnerable to the Chinese room argument (Harnad, 1989). Since the Chinese room argument is taken to show that formal computation cannot capture understanding, this criterion requires a model that essentially involves something that cannot be captured in formal computation. The point is that, if something crucial is necessarily left out when Searle in his room

engages in a formal computational simulation, then Searle in his room will not have captured crucial aspects of the model, and, therefore, the model cannot be shown to be inadequate by a Chinese room type of argument (Harnad, 1989, 1990).

Diagnosis: The Symbol Grounding Problem. Harnad's diagnosis of the Chinese room argument is that its power rests on the fact that formal computation involves only formal relationships — symbol manipulations are purely formal, and systematic relationships among symbols are also purely formal. They are all based solely on the formal "shape" of the symbols. Within the constraints of such notions of formal computation, any attempt to *define* a symbol can only relate it to other purely formal symbols. This leaves all such symbols hanging — ungrounded — and, therefore, meaningless. Searle in his room, therefore, can engage in all the formal symbol manipulations specified in any computational model, and there will still be no meaning for any of the symbols. Avoiding this regress of definitions of formal symbols in terms of formal symbols is called "the symbol grounding problem" — the problem of halting the regress (Harnad, 1990).

Causality is not Computation. Harnad proposes that any model that essentially involves causal relationships is invulnerable to the Chinese room, because any computational model of cause will be at best simulation, and will not constitute cause. Therefore, causal relations, if essential to a model, succeed in avoiding the Chinese room (Harnad, 1990, 1992a).

Transduction is Causal. More specifically, Harnad proposes a model that is "grounded" on causal transduction of sensory information. Because the regress is halted on a causal ground, mere computational simulation, so the argument goes, cannot capture crucial aspects of the model, and the Chinese room argument fails (Harnad, 1989, 1990, 1992a, 1993a).

Levels. The core idea of grounding in causal transduction is elaborated into a three-level model. The first level consists of analog transductions of sensory information (causal); this provides discrimination of one stimulus from another due to their being projected differently in the analog transduction. The second level consists of an extraction of features of the stimuli that are invariant with respect to useful categories, and serve to detect instances of those categories. This level is commonly proposed to consist of connectionist nets.

Discrimination of such categorial invariances is proposed as constituting identification of instances of the category (Harnad, 1987a, 1987b).

Categorial Perception. Harnad points out that a net making such category "identifications" will tend to exhibit a phenomenon known in human and animal perception called categorial perception. The basic idea is that senses of distance between stimuli that fall within a category is reduced relative to senses of distance between stimuli that cross category boundaries, even if their actual distances on an underlying analog stimulus dimension are equal. A classical example is sounds that fall within a phoneme category versus sounds that cross phoneme boundaries. The process of generating the identificatory classification distorts or warps the underlying analog dimensions in accordance with the categorial boundaries (Harnad, 1987b, 1993d; Harnad, Hanson, Lubin, 1991, 1994).[9]

Symbols. Harnad's third level proposes that the machine states that are generated by categorial identifications are elements in a systematically combinatorial system. In other words, these identifications will be in terms of *symbols*, capable of combinations into propositions. Higher level categories can be created via combinations of sensory level categories, as in a identification of "zebra" in terms of "horse" with "stripes," and even categories with no members, as in an identification of "unicorn" in terms of "horse" with "horn" (Harnad, 1993d).

Logic? Harnad acknowledges that some logical operators may have to be innately provided in order for this to work, and that his proposal concerning language renders all sentences as asserting category memberships, with underlying markers such as interrogative or imperative for non-declaratives. He also acknowledges that the definitions of categories that might be found in such a system will not necessarily capture necessary and sufficient conditions for any essence of

---

[9] Note, however, that *any* map from a continuous input space into a nominal output space will "warp" the input space. More generally, maps between non-homomorphic structures will induce warps in the domain. Furthermore, perceptual processing *will* generate different topologies from the input spaces because that is what such processing is for: the processing is based on the inputs, but it is "aimed" toward classification and action. The warping that is associated with categorial perception, then, should be ubiquitous, even if the output space is not categorial. Put conversely, categorial perceptual warping should be just a special case of perceptual warping in general.

Note further that if an input space is mapped exhaustively into a nominal space, and a new category is added to the nominal space, then the old category boundaries in the input space must change: if the space is originally exhaustively mapped, then changing the boundaries of old categories in that space is the only way for there to be room for a new category to be inserted into that space (Campbell, 1994).

such categories, but claims that all that is genuinely required is that the categorical classifications work sufficiently well *in fact*, and that they be modifiable if conditions are encountered in which the previous ways of categorizing no longer suffice. For example, a new environment may require new features for discriminating mushrooms from toadstools — features that were not necessary for such discrimination in the old environment (Harnad, 1993d).

Harnad does not address the problem of how combining meaningless elements, which the products of categorizations are supposed to be (Harnad, 1989, 1993d), is supposed to create meaningful "propositions" (Harnad, 1989). Nor does he address how the meaning of logical elements can be captured as an empirical category — their supposed innate origin does not touch upon this issue (Harnad, 1993c). Just what is it that is supposedly innate that is supposed to constitute a logical operator?

Satisfy The Total Turing Test. At this point we come to Harnad's primary *positive* task-criterion. Harnad argues that, because his model is grounded in causal transduction, it is not vulnerable to the Chinese room. But that, being only a negative criterion, gives no logical grounds for accepting the model as correct. The classical *computational* positive criterion has been the Turing test, but that is precisely what the Chinese room argument shows to be inadequate.

In its place, Harnad proposes the Total Turing Test. Instead of simply communicating with a machine via symbols, as in the standard Turing Test, Harnad proposes that the test be scaled up to a full robotic functionalist criterion — an equivalence of machine capabilities not only at the symbolic level, but also at the sensory and motor levels as well. A machine satisfying this test will have to have not only the sorts of discriminatory, identificatory, and symbolic capabilities outlined, but will also have to be appropriately connected to behavioral effectors (Harnad, 1991).

The Total Turing Test is a much more stringent criterion than the classical Turing test, because it requires appropriate sensory-motor functioning. That, in turn, requires some sort of causal sensory transduction, and, so the argument goes, that necessity for causal sensory transduction renders anything satisfying the Total Turing Test not vulnerable to the Chinese room argument against the Turing Test per se. Harnad claims, then, that the Total Turing Test escapes the limitations of the Turing Test.

Still No Mind, No Meaning. Even satisfying the Total Turing Test, however, still does not guarantee mind. Although Harnad sometimes writes as if Total Turing Test capability assures symbol meanings (Harnad, 1987b, 1989, 1990), in more careful moments he claims that something could pass the Total Turing Test and still be just going through the proper motions, with no mind, no subjectivity. Since meaning is ultimately a matter of subjectivity — qualia, perhaps — this yields that satisfying the Total Turing Test does not guarantee meaning either (Harnad, 1993a, 1993b, 1993d).

In fact, although the Total Turing Test puts much more constraint on any external interpretations of the symbols used by the system — more constraint than just the systematic relational constraints in standard formal symbol computational models — interpreting any symbols of such a system as having particular meanings is still just a matter of external interpretation, and does not provide any assurance that those symbols have any intrinsic meaning for the system itself. The Total Turing Test requires that any such interpretations be consistent with both the symbolic systematic constraints *and* with the sensory-motor robotic functionalism constraints, but they will nevertheless remain just external interpretations (Harnad, 1991, 1992a, 1993d; Hayes, Harnad, Perlis, Block, 1992).

Mind is Not Empirical. Furthermore, Harnad asserts, there is *no* empirical criterion that can assure mindfulness (Harnad, 1989, 1991, 1993a, 1993b). Accepting that, Harnad suggests that the Total Turing Test is the best we can do for a criterion of success in modeling "other minds" — there is no guarantee that anything satisfying the test will in fact be mindful, but there is no better test (Harnad, 1991, 1993a, 1993c).

Some Problems. At this point, we turn to some critiques of Harnad's positions and comparisons with the interactive model. In spite of a convergence between the two positions with regard to one critique of encoded symbols, Harnad's positions are in fundamental disagreement with the interactive model. They do not solve the problems of encodingism. We argue, in fact, that they are ultimately anti-naturalistic and anti-scientific.

The Infinite Regress Argument is Shared with Interactivism, But Not Much Else Is. The symbol grounding argument, with its core critique of the infinite regress of formal symbol definitions, is one of the many critiques that we propose against encodings. In that respect, then, it would appear that Harnad's proposal and interactivism might be aiming at

similar problems. The divergences beyond the overlap of this one argument, however, are deep.

Epiphenomenalism. First, Harnad's claim that no empirical test could test for mind presupposes an *epiphenomenalism* of mind. If mind were anything other than epiphenomenal, if mind had any consequences of its own, then those consequences could be tested for, and an empirical test would be possible. Furthermore, mind is not only epiphenomenal for Harnad, it is *arbitrarily* so, since he claims that no construction, even down to the level of brain functioning, could assure that the constructed system had even an epiphenomenal mind (Harnad, 1991, 1993b). Mindfulness, in Harnad's view, seems to be a strictly non-contingent epiphenomenon.

Consider the possibility that some mental property, or perhaps all mental properties, are emergent — and necessarily emergent — properties of certain sorts of system organizations. Life, for example, is an emergent of certain sorts of open systems. If so, then constructing any such system would assure the instantiation of that mental property. Suppose further that that emergent mental property had its own consequences for the rest of the system, and, therefore, for the overall functioning of the system — then those consequences could serve as tests for the existence of that mental property.

These positions are clearly those taken by the interactivist model. In seeking the emergent ontology of mental phenomena, then, interactivism constitutes a radical departure from, and disagreement with, Harnad's presuppositions. The interactivist position is one of naturalism: just as life, fire, magnetism, heat, and other once strange phenomena are now understood at least in principle as parts of the overall natural order, so also will mind be understood as emergent in the overall natural world. As such, systems with minds can be modeled, can be built, and can be tested for — in principle.

Harnad does not argue for his arbitrary epiphenomenalism. He doesn't even mention it, but, instead, presupposes it in his claims that mind is not an empirical matter, either of test or of construction. His position is contrary to the history of science, and, absent argument, seems an extremely poor bet.

An Other Minds Argument. Harnad writes in favor of his Total Turing Test by pointing out that we do in fact infer the existence of other minds on the basis of such symbolic and sensory-motor evidence — that's all the evidence that we have for inferring other minds in other people

(Harnad, 1991). Given that, Harnad urges, it would be perverse to withhold an inference of mind from any other system that presented the same evidence. He also suggests that the sense in which the Total Turing Test leaves the issue of mind unsettled is just the familiar sense in which any data underdetermines scientific theory (Harnad, 1991).

There are serious problems with this "argument" for the Total Turing Test. Here is one: The Total Turing Test criterion is a purely empirical test — in fact, a behaviorist test (Harnad, 1990). To attempt to define appropriate attribution of mind on the basis of such empiricism is a version of operational definitionalism — e.g., intelligence is whatever intelligence tests measure. There are massive in-principle problems with this sort of empiricist epistemology (Bickhard, 1992d; Bolinger, 1967; Fodor, Bever, Garrett, 1974; Hempel, 1965; Putnam, 1975, 1990, 1992; Suppe, 1977a, 1977b). For our purposes, it doesn't provide any model of the phenomena of interest — mental phenomena, in this case — and any criterion defined solely in terms of external evidence can fail to even *discriminate* the system processes of interest from other possible processes, even prior to any consideration of constituting a model of them.

A computer, for example, could engage one part of its circuitry rather than some other part, or could even execute one subroutine rather than some other subroutine, without there being any external "behavioral" evidence for the differences at all. Anything like a total external empirical test will fail to discriminate the cases. Yet there will be a fact of the matter about which circuit was engaged or which subroutine was executed, and it is not problematic in principle to test for the fact of that matter, so long as empiricist restrictions to "behavioral" data are not ideologically adhered to. The possibility that mental phenomena could be emergent in similar internal facts of the matter that are not discriminable by external data is ignored in Harnad's claims that satisfying the Total Turing Test is the best that we can hope for.

A Stronger Test — Missing Levels of Analysis. Harnad does acknowledge a stronger test than the Total Turing Test: a comparison molecule-by-molecule between the system in question and human brains. He suggests, however, that that level of stringency will not be necessary (Harnad, 1991). This point is usually put rather briefly, but it contains a serious omission that contributes to the errors concerning the Total Turing Test. By posing a dichotomous choice between molecule-by-molecule comparisons and external Total Turing Test comparisons, Harnad

indirectly appeals to the functional intuitions that most of his readers share. Very few researchers in Artificial Intelligence or Cognitive Science would hold that molecules are the relevant level of analysis, but that seems to leave the Total Turing Test as the alternative.

Note, however, that standard functionalism is not to be found in this dichotomy. The distinction between this subroutine rather than that subroutine might not make any difference at the level of external behavioral data, but it will be a fact of the matter at a functional level of analysis, without having to examine the "molecules" of the computer. Any version of functionalism will constitute a counterexample to Harnad's dichotomization.

Harnad might wish to counter that the Chinese room has already shown the inadequacies of functionalism, but that point, even if accepted, does not address the possibility of other levels of analysis situated above molecules but more internal to the system than behavioral data. Furthermore, acceptance of the Chinese room argument is far from universal. And still further, this rejoinder on Harnad's behalf would equate functionalism with formal computationalism, an assimilation that is also not universally acceptable. In fact, there is an argument that the basic insights and intuitions of functionalism cannot be made good except within the interactivist approach (Bickhard & Richie, 1983; Bickhard, 1993a), which is *not* a computationalist approach.

Empiricist Epistemology — Empiricist Semantics. In discussing his own approach to language, Harnad acknowledges that the empiricism he espouses has been severely criticized (Harnad, 1987b, 1993c). His response, however, is that such an empiricist approach has not really been tested (Harnad, 1992a), and to just assume the problems away (Harnad, 1993c). The problems, however, are problems in principle, and no finite number of empirical tests can discover an in-principle flaw. Harnad's appeal to ignore these critiques and simply proceed with "testing" is simply asking for a license to ignore reason — and history (Coffa, 1991; Suppe, 1977a). It does not provide rational grounds for accepting or even pursuing such empiricist errors.

In fact, such behaviorist empiricism *has* been "tested," and has been found fatally wanting. The points about some facts of the matter internal to a computer being externally non-discriminable provide unbounded classes of counterexamples. That was one of the primary lessons that computers provided to the recognition of the inadequacies of behaviorism some thirty or forty years ago. These fundamental

inadequacies of empiricism, then, show up both in Harnad's proposals about language, and in his proposals about the Total Turing Test being all that we can hope for in attempting to study mind.

Infer or Discriminate? There is also an inconsistency in Harnad's positions concerning categorization and concerning the Total Turing Test. Harnad points out that we discriminate categories using whatever works, and without necessarily knowing or invoking defining or essential conditions. That is why, among other consequences, that we on occasion discover that we have to sharpen our discriminations — we discover circumstances in which discriminations no longer suffice (insofar as discriminations are differentiations, this point is convergent with the interactivist model). But, in his discussion of the Total Turing Test, and in his defense of that Test on the basis of a comparison with the issue of other minds, Harnad alleges that we *infer* other minds on the basis of the symbolic and behavioral external evidence — the behaviorist evidence — provided by other bodies (Harnad, 1991, 1992b). Harnad doesn't focus on the world "infer," but it is an interestingly strong word compared to the merely context dependent empirically adequate discriminations that we are supposed to engage in for most, or all the rest, of our categories.

If we re-consider Harnad's other minds discussion from within the framework of *context dependent discriminations*, rather than *inferences*, we note that we do *not* infer other minds, on the basis of anything. Instead, we discriminate entities which we treat in ways that presuppose mind from other entities that we treat in ways that do not presuppose mind. And, if we found that our discriminations were no longer working — say, in some new environment, populated perhaps by mindless robots satisfying Harnad's Total Turing Test — we might well sharpen our discriminations, perhaps even including some internal functional criteria. At least we would seek additional criteria, and the problem of *which* criteria would be most satisfactory *is* the problem of how to model mind. If empirical criteria were available that worked, we might adopt them, while if, say, functional — internal — criteria were required, we might adopt them. In general, of course, we tend to use criteria that are readily available, even if they are known to be fallible with respect to more stringent but also more difficult criteria. So, we might discriminate mindfulness on the basis of easily obtained evidence even after we have discovered that those criteria do not work in certain circumstances — so long as the cost of failure in discrimination is not too high.

Currently, of course, very coarse discriminations suffice (so far as we know) to pick out mindful entities from nonmindful entities. Harnad's appeal to our current "inferences" of other minds on the basis of behavioral evidence, then, is simply question begging. What we currently do to discriminate mindful from nonmindful has little bearing on what mind might be, and similarly has little bearing on what we *might* use for such discriminations in circumstances in which more careful discriminations were advisable. The Total Turing Test, correspondingly, has little to do with what we would or should take as criterial for mind or meaning or intentionality. Those are questions of science and philosophy, not of empiricist epistemology.

Mere Reverse Engineering. In accordance with his presupposition that subjectivity is epiphenomenal, Harnad — by encompassing all issues of genuine meaning, representational content, qualia, and so on into subjectivity — can then dismiss addressing those issues on the basis of his claims that 1) they are of no empirical consequence, and 2) therefore they cannot be scientifically investigated (Harnad, 1991). By so dismissing the difficult questions, he is able to claim that satisfying the Total Turing Test is the best that can be hoped for, and that satisfying the Total Turing Test is a matter of *reverse engineering*, and *not* a matter of basic science (Harnad, 1993a). Furthermore, this is reverse engineering that need not try to address fundamental issues such as content, and so on, since those are not empirical matters anyway (in spite of some contrary hints in Harnad, 1989). On such a view, a table could be reverse engineered without having to address such basic science issues as valence, atomic bonding, intermolecular forces, and so on. Harnad's alleged scientific stance in fact amounts to a hand-waving dismissal of all of the most important scientific questions, with non-argued claims that those questions are not empirically investigatible anyway. Harnad's proposals amount to an unargued rejection of the fundamental naturalism that has guided science for centuries. Again, that does not seem like a good bet.

Self Insulation. The deepest difference between interactivism and Harnad's positions, then, is that the fundamental problems that interactivism attempts to address are presupposed by Harnad as being not scientifically addressable. In so doing, he insulates his own positions from criticisms concerning their multiple failures to address those fundamental issues. Interactivism does attempt to address such issues,

and, for at least some of them, such as representational content, claims to have a model.

Empiricism Writ Large. All of these points —

- problems with logic,
- problems with language,
- the epiphenomenality of mind,
- the mis-use of the other minds issue,
- the dichotomization between a Total Turing Test and molecule-by-molecule comparisons,
- empiricist semantics,
- rejection of in-principle arguments,
- the mere reverse engineering claim, and
- self-insulation against foundational problems —

are simply rehearsals of familiar failures of empiricist epistemologies specialized to the discussion at hand. Harnad's proposals, and his "defenses" of his proposals, deeply presuppose an empiricism that cannot be sustained against any sort of careful considerations.

Additional Problems. There are still further problems with Harnad's positions. We take a look at four of them:

- A circularity in Harnad's argument concerning the invulnerability of transduction to a Chinese room style argument;
- An odd assumption that issues of meaning apply only to formal symbols;
- An inadequate notion of learning, and;
- Both a claim and a disclaimer of the relevance of Harnad's model to issues of intentionality.

What's Special About Transduction? Harnad's argument that causal transduction is not vulnerable to a Chinese room argument (Harnad, 1993a, 1993b) seems to be based on a circularity, manifesting an underlying equivocation. We outline the circularity first. In the Chinese Room argument, the crucial assumption is that Searle can be doing *everything* that a formal system he is implementing would be doing — receiving all the symbol inputs, manipulating in accordance with all the rules, emitting all the symbol outputs, and so on — and *still* there would be no understanding of Chinese going on. The argument turns on the claim that the formal system can be fully implemented without implementing *understanding.*

Harnad argues that, in contrast, Searle cannot fully implement a causal transducer without thereby implementing actual *seeing*. This argument, then, turns on transduction and *seeing* instead of on formal systems and *understanding*, and Harnad's claim is that a Chinese room style of argument fails in this case — the parallel does not go through. Harnad claims that, if Searle attempts to implement a *causal transducer*, then either:

- Searle — *himself* — *is* the transducer, in which case Searle *sees*, and the implementation is not merely a simulation, but *is* the actual phenomenon of interest — "seeing," or
- Searle himself is *not* the transducer and instead he simply receives the outputs of transducers. In this case Searle's attempted rendition of the basic transducer model has left something crucial out, and the failure of "seeing" to occur is simply due to Searle's failure to capture the critical aspects of the transducer model.

In particular, causality (transduction or otherwise) cannot be captured via mere formal simulation — the causality has to actually occur. In contrast, a computational implementation (or simulation) of a computation, so the argument goes, is itself a computation — and, in fact, can be exactly the same computation, to all formal criteria, of the computation being "simulated." This isn't so for causal phenomena: a computational simulation of a thunder storm does not get anything wet. So, unlike computational models, causal transduction models are not vulnerable to a Chinese room argument because they cannot be captured merely by causal "simulation," and they cannot be *implemented* without the necessary causal processes actually taking place.

Harnad's argument, however, rests on the claim that, if Searle is not himself the transducer — in which case Searle would himself be "seeing" — then Searle is not capturing the transducer model (Harnad, 1993a, 1993b). The issue, then, is what does a transducer actually do, and what would Searle have to do to capture a transducer model, to implement it. *The supposition is that a machine that was **really** engaged in such transduction would be seeing*. Therefore, since Searle isn't seeing, Searle isn't capturing the transduction model. That is, if Searle isn't actually seeing, then Searle is not implementing the model.

By shifting the structure of this argument back into the original Chinese room framework, it becomes evident that there might be something wrong with it. Suppose that one were to argue against the

Chinese room argument that Searle isn't capturing the original model because, *unlike* a machine that was actually engaged in the formal model activities, *Searle isn't understanding Chinese*:

- If the machine would be understanding Chinese, and if Searle isn't understanding Chinese, then Searle is failing to capture the activities of the machine.
- So, if the machine would be seeing, and if Searle isn't seeing, then Searle is failing to capture the activities of the machine.

This position clearly begs the question. Whether or not a machine manipulating all those Chinese symbols in appropriate ways would be *in fact* understanding them is the issue at hand in the first place. Similarly, whether a machine engaged in all those transductions, or causally receiving all the outputs of those transductions, would *in fact* be seeing is the issue at hand in Harnad's transduction claims. To claim that Searle fails to capture what the machine would be doing so long as Searle isn't "seeing" is a simple circularity. There may well be important properties of causal transduction, but Harnad's *argument* does not succeed in discriminating them, and certainly not in modeling or explaining them.

If only the *causal* properties of transduction are supposedly at issue — and not any alleged or presupposed intentional properties that require Searle the *epistemic* agent (not any properties that require Searle to be *seeing*) — then Harnad provides *no* account of why Searle being a *causal* transducer *without* any seeing going on (without any understanding going on) would not count as a counterexample. Searle could, for example, transduce sun exposure into redness of sunburn, or into damage to rods and cones, or into severity of squint, and so on. None of these involve "seeing," but, then, neither do any other forms of causal transduction that anyone has ever heard of — photocells, speedometers, cameras, and so on. Couldn't Searle "implement" a photocell that opens a door without Searle actually seeing?

Of course, none of these is involved in a Total-Turing-Test-competent robot, but the relevance of that criterion has already been shown to be questionable. In any case, the core of Harnad's claim for the special invulnerability of causal transduction to a Chinese room argument is *not* based on Total Turing Test competence per se, but rather on the alleged necessity for Searle to be seeing in order for Searle to be capturing what the transduction machine would be doing. And that argument is circular independent of any issues about the Total Turing Test — that argument assumes that the machine would in fact be seeing.

At this point, the underlying equivocation is clear: Harnad's argument that Searle would have to be *seeing* in order to implement causal transduction begs the question because he smuggles cognition — seeing — into his usage of the notion of transduction, when transduction is *just* a causal relation. By equivocating on these two usages of transduction — cognitive and causal — we get Harnad's argument.

Do Issues of Meaning Apply Only to Formal Symbols? There is a peculiarity of Harnad's discussion of his infinite regress argument that relates to this issue about transduction. Harnad claims that the infinite regress argument applies *only* to symbols (Harnad, 1990, 1993b), and he claims that symbols are symbols *only* by virtue of their being elements in a systematic combinatorial organization, such as the systematicity of combining words into sentences (Harnad, 1992a, 1993d). He claims that the infinite regress argument does not apply to the products of analog transductions or neural nets, for example, because those products are not elements of systematic symbol systems (Harnad, 1992a, 1993a, 1993b), and, therefore, that the grounding problem does not apply to such transductions. The presuppositions underlying these claims are not clear, but one that could lie behind such a position would be a presupposition that the infinite regress of definitions can only occur within a systematic symbol system, since that regress of definitions requires that there be such a resource of symbols in the first place. Because a transduction product is not an element in such a system, no such regress of definitions could exist, and, therefore, no such problem arises.

Note, however, that the only reason for invoking anything like the infinite regress of definitions within a symbol system is an attempt to provide meaning to the symbols (Harnad, 1989, 1990, 1992a, 1993d). The significance of the regress is not that the regress per se is possible, it is that *even with* such a regress — *even with* such a resource for definitions — there will *still* be no meaning for any of the symbols. The product of a transduction, then, ought to be in even *more* trouble than a (systematic) symbol, according to Harnad's positions here, since it does not even have the resource of systematic regresses of definitions to provide any meaning, as inadequate as that resource ultimately proves to be. Does an attempt to provide meaning via definition have to be in infinite regress in order to fail to provide meaning? For a transducer or a net output, the definitional regress can't even *begin*.

The problem of semantics for symbols is not created by the systematicity of a symbol system. It is not created by the possibility of an

infinite regress of definitions. The problem of semantics for symbols, rather, *fails to be solvable* even if such a regress of definitions *is* possible. If such a regress is not possible because the transduction or net product does not belong to a systematic symbol system in the first place, that contributes nothing to the solution of the problem of semantics.

Harnad has focused much too narrowly on the infinite regress of definitions as somehow creating the semantics problem, thereby failing to recognize the true scope of the semantics problem. That narrowness of focus on the source of the problem of semantics, in turn, has been shored up by an artificial restriction — a supplemental narrowness of focus — of the notion of symbol to an element of a systematic symbol system (Vera & Simon, 1994). With that restriction, the *symbol grounding problem,* in Harnad's usage, only applies to symbols in symbol systems because only such elements are symbols at all and only such elements have the resources for infinite regresses of definitions.

In this view, the problem of grounding — and, therefore, the problem of semantics — cannot apply to transduction or net products that do not belong to such systems because, again in Harnad's usage, such products are not symbols at all if they are not elements in symbol systems, so the *symbol* grounding problem cannot apply to them because they are not symbols in the first place (and, therefore, cannot have infinite regresses of definitions within such systems). Finally, if the symbol grounding problem does not apply, then the problem of semantics does not apply — in Harnad's usages. So, if transduction or net products are not symbols at all (because they are not elements of symbol systems), then they are not subject to the infinite regress of definitions problem, and, therefore, they are not subject to the symbol grounding problem — *and*, therefore, they are not subject to a problem of semantics.

The backbone of this position seems to be: If 1) the problem of semantics is construed as being created by the possibility of an infinite regress of definitions; and 2) such a regress is construed as being possible only within a symbol *system*, then 3) the problem of semantics would exist only for such elements of symbol systems. Premise 1), however, is simply false, and the rest of the confusing circularities of definition of "symbol" and "symbol grounding" are in support of that initial false assumption. Why, independent of such circularities and arbitrary restrictions of definition, would the problem of semantics, of representational meaning, be restricted exclusively to elements in a symbol *system*?

There are also still further questions. Why for example, is the product of a transducer — a point in an analog space, for example — not just as systematic in its own way, with respect to the organization of that space, as a typical formal symbol is with respect to the organization of *its* space of dynamic possibilities. It can also be questioned why the symbols that might be used in an attempt to give meaning to, say, a transducer or net product, have to be in a systematic symbol system *together with* that element that is to be defined. That is, why wouldn't *any* attempt to define the meaning of a net product be just as subject to an infinite regress of definitions problem, even though that infinite regress occurred within a symbol system that does not include the net product itself? How else is that transducer or net product to acquire any meaning?

At this point, Harnad's answer is likely to be in terms of the causal and/or analog relations between transduction or net products and that which they have transduced. Harnad's response in terms of analog transduction, however, purports to answer a question about a *representational* relationship in terms of a strictly causal or *factual* relationship. At this point, in other words, we return to Harnad's equivocation between causal and cognitive usages of the term "transduction." As we have seen, such causal or factual relations do not constitute representational relations, and therefore, do not provide any solution to the problem of meaning for transduction or net products. Do visual transductions, for example, represent light patterns, or retinal stimulations, or objects and surfaces, or electron orbitals in objects and surfaces, and so on? And how could a system with such visual transduction inputs have any representational information about which of such possibilities those inputs are supposed to represent (Bickhard & Richie, 1983)? And so on. We will not pursue these additional issues any further here.

In sum, transducer and net products are just as much subject to a problem of semantics as are symbols in symbol systems. The infinite regress problem does not create that problem, it merely fails to solve it. So, the symbol grounding problem either *applies* to transducer and net products, if the symbol grounding problem is identified with the problem of semantics, or the symbol grounding problem is *irrelevant* to the problem of semantics, if it is identified with the infinite regress problem. Harnad equivocates between these two usages of "the symbol grounding problem," thereby presupposing that the problem of semantics is *equivalent* to the infinite regress problem. In either case, the causal or

analog properties of transduction and net products do not solve the problem of semantics. Therefore, transduction and net products have not been shown by Harnad to have any advantage over systematic symbols (including *transduced* systematic symbols; Fodor & Pylyshyn, 1981; Bickhard & Richie, 1983) with respect to issues of semantics and meaning.

Learning? Christiansen & Chater (1992) question how a system could solve the problem of error that afflicts *all* causal correspondence models of representation: How can a correlation be wrong? In response, Harnad claims that there is in fact no problem about learning in his model (Harnad, 1993d). He illustrates with a story about a system learning to differentiate mushrooms from toadstools on the basis of whether or not eating pieces made the system sick.

There are in fact at least two problems with this response. First, not all learning is based on innate error criteria, such as nausea or pain. Harnad, in fact, seems unable in principle to be able to account for any but supervised or tutored learning — such as the tutoring for back-propagation in neural nets — in which the tutor already "knows" the right answer. In this regard, innate error conditions such as nausea or pain are simply evolutionary learnings of what constitutes error — the organism is being tutored by the innate error criteria. Such learning certainly occurs, but it cannot account for all learning, and it cannot account for the learning of new error conditions. Mathematics, for example, involves ever more sophisticated and complex error conditions (e.g., overlooking a presupposition of the axiom of choice in a proof). These are not innate.

The second problem is even deeper. What the system in the mushroom-toadstool example is actually learning is that one internal condition of the system indicates "appropriate for eating" and another internal condition indicates "not appropriate for eating." It happens that, in this fable, conditions for eating happen, as a factual matter, to be causally induced by mushrooms, and conditions for not eating happen, factually, to be causally induced by toadstools. But the system learns nothing about mushrooms or toadstools in this story, only about how to discriminate eating conditions from non-eating conditions.

In particular, it is not the correspondences or correlations with mushrooms or toadstools that are learned at all. Those correspondences remain at a purely causal level, and serve *only* to induce the internal indications for eating or not eating. There is no learning of any concepts of or meanings of or references to mushrooms or toadstools. The sort of

learning that Harnad outlines can occur, but what is being learned is an indication of a potentiality for engaging in certain further actions — eating, for example — associated with internal functional expectations of particular internal consequences of those actions — satisfaction or nausea, for example.

This is, in fact, an example of *interactive* representation — representation of *action-to-internal consequence* relations, contingent on ***contentless*** *differentiations* of the environment (contentless differentiations, for example, between mushroom-environments and toadstool-environments). Such indications of potential internal consequences *can* be in error, and *can* be found to be in error, by the system itself. So systems *can* learn such indications, and those indications can be correct or false, and falsified, for the system itself. Still further, *any* such internal functional interactive indication can be similarly found to be in error by the system itself — innate error signals are not required.

The fact that the differentiations that properly yield "eat" indications are factually correspondent to mushrooms, and those that properly yield "don't eat" indications are factually correspondent to toadstools, explains (to an external observer, for example) *why* such indications, based on such differentiations, are useful to the system. But they do *not* constitute any *system* knowledge or information about mushrooms or toadstools per se. They do not constitute meanings about the external ends of the causal analog transductions. They do not constitute meaningful "groundings" of the symbols that Harnad labels as "mushroom" or "toadstool."

In this fable, then, Harnad has grabbed a small piece of what we advocate as the basic model of representation, but he has misconstrued it. He has suggested that the system learns about mushrooms and toadstools, but his story at best supports the conclusion that the system has learned a way to distinguish "eating" from "not eating" situations. For a way to fill out this model of representation that avoids that error, we suggest the interactive model.

What Happened to Intentionality? Harnad originally relates the symbol grounding problem to the problem of intentionality (Harnad, 1989, 1990), so to claim that analog transducer outputs are not subject to the symbol grounding problem ought to entail that they are not subject to the problem of intentionality. But they are, and, even according to Harnad, they are (Harnad, 1990, 1993a, 1993d). So, it would seem that

either the symbol grounding problem is *not* related to the problem of intentionality, or that analog transducer outputs *are* subject to that problem. In either case, there appears to be a contradiction.

It is peculiar that Harnad insists that transducer products are not vulnerable to the infinite regress argument, but then turns around and claims that, even with Total Turing Test capability, complete with such (analog) transducers, a system might *still* have no mind, no content, no meaning. Even with Total Turing Test capability, meaning will still be an issue of external interpretation, not intrinsic meaning for the machine itself (Harnad, 1993d). If that isn't being subject to a "symbol grounding problem," then it is not clear what is.

If Harnad's "symbol grounding" does not address issues of intentionality, and if "symbol grounding" is accomplished by causal analog transduction (plus "categorization" and systematicity), then "symbol grounding" is reduced to merely a name for causal analog transduction (plus, etc.). Because "symbol grounding" does not address issues of intentionality, such as meaning (Harnad, 1990, 1993a, 1993c, 1993d), it is not clear what issues it *is* addressing, or is relevant to. Harnad writes as if he *is* addressing basic issues (Harnad, 1989, 1990), but then retracts such suggestions — claiming that he isn't addressing them after all, because they are not empirical issues. In consequence, it is not clear *what* Harnad thinks "symbol grounding" is relevant to. At a minimum, it *is* clear that none of Harnad's model, in his own terms, even addresses, much less solves, the basic problems that interactivism confronts.

Epiphenomenalism versus Naturalism. Overall, although Harnad and interactivism partially share an appreciation of an infinite regress argument, there is little else they have in common. Harnad's position makes the assumption that causal analog transducers — hooked into categorizing nets and systematic symbols — will somehow generate meanings. Or, at least, if part of a Total-Turing-Test-competent system, they will generate meanings. Or at least (even if not) this is the best that we can hope for. The interactivist approach proceeds on assumptions contrary to these on all levels.

Interactivism is a *naturalistic* position, and attempts to model the *emergent* nature of *representation* — genuine representation — in system organization. If the interactive model is correct, and if a system were constructed with such organization, then that system *would* have representational content — *even if it didn't satisfy a Total Turing Test*.

The Cartesian Gulf, Again. Note that requiring a Total Turing Test competence (or any other kind of fully-human competence criterion) before being willing to grant *any* mental property commits the Cartesian-gulf error of assuming that all mental properties are a necessary whole, and that evolution, for example, did not generate the emergence of some mental properties before, and without the simultaneous presence of, other mental properties. (See the discussion of Searle above.) Neither fish nor snakes nor frogs nor dogs nor monkeys could pass the Total Turing Test. It does not follow that they have no representational contents.

Empiricism — Behaviorism. Interactivism is concerned with the ontology of its subject matter — representation. In this focus on ontology, on the nature of the phenomena, it shares the goals of physics, chemistry, biology, and virtually all other sciences. Only psychology and some branches of sociology ever swallowed the poison of behaviorist empiricism as a conception of good science (Bickhard, 1992d). Behaviorist empiricism is a fatally flawed conception of *any* science, and is a self-contradiction as an approach to matters of mind. Harnad seems to be struggling within the hall of mirrors of that approach (Harnad, 1990, 1993a, 1993c).

Representational Content: The Real Issue. Most fundamentally, Harnad has *no* model of representational content, and cannot even attempt one within his epiphenomenal presuppositions. He has, on his own terms, some criteria — e.g., analog causal transduction — that he thinks might be necessary for a model attempting to solve his "symbol grounding" problem, and an empirical criterion — Total Turing Test capacity — that, although it too will ultimately *not* address issues of meaning, is, so Harnad claims, the best we can hope for. Considered from within the interactivist framework, transduction — analog or otherwise — does nothing to solve the encodingism problems, and cannot provide a model of representational content. Epiphenomenalism is an anti-naturalism; it is anti-science. Interactivism, in contrast, *does* propose a naturalist model of representational content, of the ontology of representational emergence.

Radu Bogdan

Bogdan (1988a, 1988b, 1989) proposes a naturalistic ontology of representation, though he uses "representation" in a more restricted sense than we are here. Our representation is roughly his "semantic information" in scope, though certainly not in definition (see below). He is concerned with the embedding of representation in the activities of

actual organisms in an actual world; in this there is certainly a parallel with the interactivist approach. He is also cognizant of a critical importance of goals to representation — even more of a parallel. Beyond this, however, divergences exponentiate.

Bogdan's representation (semantic information) is basically constituted as appropriately registered (from the environment) and organized internal states with distal sources as their encoding contents. Those representational contents are carried by those internal states in the sense that the teleology of the system is *explainable* only in terms of such connections between the internal states and the distal sources: e.g., an animal's behavior with respect to (internal representations of) prey or predator. At times, Bogdan puts strong emphasis on the necessity for information to be appropriately registered and organized — constrained — in order for it to count as semantic, but the point here seems to be that the registration and organization of input must be such that it can be explained only in terms of presumed epistemic connections with distal sources. In other words, it is subsumed in the explanatory teleological dependence.

Bogdan's naturalism is clear here. But teleology and goal-directedness are of critical importance for Bogdan only for the explanation of the *existence* of representation — representation exists because it is so functional for teleology — and there is no connection for Bogdan between teleology and the constitution, the *ontology*, of representation as there is in the interactivist model. He does not recognize that system goal-directedness is necessary for representational "aboutness" to emerge, is necessary for there to be any possibility of a representation being right or wrong. He even clearly states that his model would allow for completely passive representors (semantic information systems) with no outputs (so long as things were "appropriately" registered and organized and constrained), something quite impossible within interactivism.

It turns out, furthermore, that the teleology that is so crucial to explaining the existence of representation for Bogdan can just as well be a teleology of a *designer* of the system as it can be a teleology of a system itself. For example, an electronic-eye door opening system satisfies this criterion since the designer of the system had people and other moving objects in mind when designing the photocell-to-behavior relationships. (It is presumably in terms of a designer or explainer "teleology" that a completely passive system could be semantic. It's hard to imagine an

intrinsically passive teleological *agent*.) In this move to a designer teleology, a move little explicated or defended by Bogdan, he has shifted from a naturalistically motivated encoding model to an observer idealism encoding model. As a result, whether a system is representational or not depends on the observers' design for or explanation of the system.

Bogdan reserves the word "representation" for explicitly encoded "semantic information" that can be internally operated upon by the system. He restricts "representation" to data structures and explicit symbols — either analog or digital — in the classic computer model sense. We see no reason to follow in this arbitrary restriction, as it simply confuses the issues. It is a symptom of the non-naturalistic computer metaphor myopia that dominates contemporary Artificial Intelligence and Cognitive Science, and obscures Bogdan's acknowledged naturalistic partial insights.

The strain that is introduced by this myopia is manifested in Bogdan's naming the most general and primitive form of representation — the most general ontology that has any "aboutness" — "semantic information," in spite of the fact that it has nothing to do with language, while reserving "representation" for data structures, with all of their language-like properties. The emergence of representational "aboutness" out of non-representational phenomena is the form of emergence that is most critical and difficult to account for , and it is not "semantic" except in the encoding view that makes language simply a re-encoding of cognitive encodings. In the interactive model, in fact, language itself is not directly representational at all, but is rather a system of operators on representations — it is a new level and kind of emergence from representation per se. Even in the encoding view, however, the emergence of an ontology of "aboutness" is still more critical than the emergence of explicitly encoded, manipulable *versions* of an ontology of "aboutness." In part, this is merely an issue of the stipulative "semantics" of the lexical items involved, but more deeply it is a manifestation of confusion concerning the location of the most fundamental issues.

A Metaphysical Commitment to Encodingism. Bogdan is already committed to an encoding view of representation in the prior metaphysics that he brings to bear on the questions at issue — a metaphysics that he also shares with many others. Roughly, he proposes that everything, all actual *tokens*, are instantiated materially, and that different levels of abstraction, of *types*, correspond to different levels and forms of constraint that are taken into account in *defining* those types.

The necessary advertence to some form of *definer* in understanding ontological types already requires an observer or definer or designer idealism when applied to representation, but the deeper problem is in the metaphysics of abstractive types and tokens.

If representation is construed as some sort of abstracted *type*, with various materially instantiated tokens, then the only differentiating characteristics that are available for distinguishing representational types from other types are precisely the *representational characteristics* that are supposedly to be accounted for in the ontology in the first place. The representational types must be characterized as types with defining representational properties, of which the most fundamental is representational "aboutness." To define a type in those terms, however, creates two problems: 1) it is either *circular* in that it is "aboutness" that is to be accounted for in the first place or *idealistic* in that the aboutness is simply referred to the definer or user of the type, and 2) it is intrinsically encodingist — any representation defined in terms of its representational content *is* an encoding.

Process Ontology. The metaphysics of abstractive types and tokens, however, is not sufficient for much of the world, and certainly not for representation. For one large class of counterexamples, consider the ontologies of process. There are closed system stable processes, such as atoms and molecules, and open system stable processes, such as flames and life, not to mention the many non-stable forms and patterns of process (Bickhard, 1993a; Bickhard & D. Campbell, in preparation). To be sure, all of these involve instantiation in material terms, *but not just in terms of forms of material types*. A flame is not just the molecules and atoms constituting it — the same material engaged in different interactions would not be a flame, and differing material substrates are in fact involved at each moment of the flame. Further, not only are there no flame substances or substance types, neither are there any discrete flame *states.* A flame is not a sequence of transitions from one state to another, or even from a single flame state to itself: states are simply the wrong ontology for flames. Processes are neither substances nor substance types nor states nor state transitions. Even if the state is taken in a mathematical state space sense in which the possible states form a continuous space, still no single state can constitute "flame" — the state space is a mathematical abstraction in which, at best, certain emergent properties of suitable *trajectories* in the space can represent flames — the mathematical property of continuity, and the ontological property of

"ongoingness," of process, cannot be defined in terms of isolated single points. *Ontologically*, flames require duration. We do not intend to develop a full process ontology here, only to point out that neither types and tokens, nor states, are a sufficient ontology for the world (Bickhard, 1993a).

Functional Matters. The currently relevant point, of course, is that these are not adequate ontologies for representation either. In the interactivist model, representation is a functional matter, and function is itself an ontology emergent from process (we skip several layers of emergence here; see Bickhard, 1993a), so the interactivist notion of representation cannot be captured in a type-token or state metaphysics. The only available approach to representation within Bogdan's metaphysics is encodingism, thus the intrinsic commitment to encodingism even before representation per se is addressed. An attempt to capture the *functional* aspects of representation in terms of ontological tokens not only forces an encodingism, it also highly motivates, if not forces, an *idealism* since "function" cannot be defined in terms of materially instantiated types either — it is a relational ontology, not a substance ontology. The functional properties, then, get pushed into the *definitions of the ascribed types* instead of being capturable in the ontologies, including relational ontologies, of the phenomena.

Substance Categories. It is possible, of course, to apply a category system of types and tokens to any phenomena whatsoever, including that of representation, even as interactively understood. But such an application is not generally confused with a metaphysics or ontology — it is a system of classification whose relationships to underlying ontologies and issues of emergence out of lower level ontologies is unaddressed in the categorizations per se. There is also a sense in which we would agree that all ontologies are in some sense materially instantiated, including process and relational and functional and representational ontologies. But Bogdan only allows for one form of movement to more abstract forms of types — abstraction away from the particulars of the ultimate material instantiations. This is what commits him not only to a type-token categorization system, but to the more particular *substance metaphysics*. Everything is just substance, though viewed in terms of more and more abstract type definitions. This will not handle *relations*, for example, however much it may be that relations are materially instantiated. An instance of "aboveness," for example — such as *a book above a table* — cannot be construed as an abstraction *away*

from the properties of the "things" that instantiate the relation (Olson, 1987). More generally, any properties of *patterning*, of substance (should any such thing as substance exist — see Bickhard, 1992a, 1993a, in preparation-b; Bickhard & Christopher, in press) *or* process, cannot be captured in a strictly abstractive type-token metaphysics. This includes process, function, and representation.

Primary Property Metaphysics. More generally, in only considering *abstraction from particulars* as a generator of new types, Bogdan eschews all issues of higher level pattern: of relationship, of process, of function, and so on. He implicitly commits to a primary-property-only metaphysics — no secondary or relational properties allowed, except as properties to be explained away. There are no genuine ontological emergences at all in this view, only more and more abstract types to be instantiated at the purely material substance level (whatever that is — atoms? protons, neutrons, and electrons? quarks? strings? preons? vacuum topologies?).

On a still more general level of comment, this discussion presents an instance of yet another way in which a commitment to encodingism can be implicit in what appear to be distant considerations and presuppositions. Encodingisms are far more prevalent than just the usages of the terms "encoding" or "symbol."

Bill Clancey

In recent papers (1989, 1991, 1992a), Bill Clancey has been developing a critique of standard Artificial Intelligence conceptions of representation and proposing his own interpretation of a position called "situated cognition." He has argued that AI knowledge bases, or "knowledge level" descriptions (in Newell's (1982) phrase), should be seen as observer constructed models of patterns of agent-environment interactions, rather than as mechanisms or structures internal to an agent. This implies that "knowledge engineering" must be recognized as a distinct discipline from cognitive modeling, and that each must be pursued on its own terms. We trace his argument briefly and consider it from the perspective of interactivism and from our critique of encodingism.

Clancey makes a two-faceted argument: 1) Knowledge level descriptions are properly and productively seen as observer descriptions of systems-in-the-world, i.e., of recurrent patterns of agent-environment interactions, and 2) Knowledge level descriptions are not isomorphic to

structures or mechanisms that are internal to an agent — by implication, the "representations" of AI systems are different in kind from the "representations" of human beings.

Point 1) clearly is true — i.e., the structures in AI programs that are thought of as *the rules for* medical diagnosis or computer configuration certainly are constructed by observers (e.g., the knowledge engineer) about patterns of interactions between agents and environments (e.g., one or more physician's statements about how they relate test results and patient symptoms to diagnoses). It is just as clear that they are intended to be *models*, i.e., to enable prediction and simulation of a particular system-in-the-world. In arguing that it is productive to view AI programs as models of systems-in-the-world, rather than as *cognitive* models, in part Clancey simply is giving theoretical support to what already tends to happen in AI. That is, practical projects abandon the programmatic claims of Artificial Intelligence to better achieve the pragmatic goal of designing interesting and useful systems. However, there are additional consequences of his argument, even for practical projects. For example, "knowledge engineering" cannot be seen as "transfer" or "transmission" of knowledge from the expert's head and judged by the fidelity of the transfer. Rather, it is a modeling process in which expert and knowledge engineer, as well as users and other interested parties actively work together to create appropriate models. This "worker centered knowledge engineering" (Clancey, 1992b) marries the technology of Artificial Intelligence with methodologies derived from sociological approaches to design (Norman & Draper, 1986; Greenbaum & Kyng, 1991).

Clancey supports his second claim — that knowledge level descriptions are not isomorphic to structures or mechanisms internal to an agent — in a number of ways:

- The empty symbol problem — for a program, "every problem is like assembling a puzzle with the picture-side facing down" (Clancey, 1991, p. 376). The symbols are about nothing. Clancey argues that a map — a set of correspondences — cannot provide content for symbols. Rather, content must be a property of ongoing behavior;
- The notations used in computer programs are just like any other writings in that they are liable to contextual interpretation and reinterpretation by humans. For example, Clancey notes that, as MYCIN evolved, its

designers gradually reinterpreted various symbols that appeared in its rules, thus changing the interpretation of MYCIN's "knowledge" *for the designers*, but not for MYCIN;

- A program's stock of symbols and their interpretations is supplied and fixed by the program's designers; there is no content for the program itself, and no way for the program to generate new content;
- More generally, a description generated from observable behavior (i.e., of agent-environment interactions) is not the same as the internal mechanisms that produce the behavior. To think so is a category mistake, in Ryle's (1949) sense.

From an interactivist perspective, everything that Clancey says here is true of encoded "representations," and since we have argued that standard conceptions of representation in Artificial Intelligence do construe representation as encoding, we are in full agreement with his critique. He further argues that a crucial question that analyses of knowledge must answer is how a robot can create new ways of seeing the world *for itself*, rather than being limited to the structures with which a designer has supplied it. We have argued that it is precisely the notion of knowledge *for an agent* that theories of representation need to explicate, that encodings cannot do this job, and that interactive representation can.

A Potential Problem. However, we find a potential serious problem arising from Clancey's diagnosis of knowledge as observer ascriptions about an agent. While this is an excellent design stance, it is dangerous when taken as a part of an explanation of natural intelligence. The danger here is the specter of idealism, and the resolution to that danger provides its own perspicacious perspective on the issues in this area. Consider several statements taken from Clancey (1991):

> I claim that the essential matter is not "how does the architecture support knowledge," but rather, why would we ascribe knowledge to the behavior produced by such an architecture? (p. 27)
>
> Claiming that knowledge of a certain type is a possible ascription that could be made about a given architecture requires specification of the world, tasks, and *observers* in which the architecture is embedded. (p. 28 — emphasis added)

As noted, an analysis of knowledge *for an agent*, i.e., an architecture that supports knowledge, is just what we need for a naturalistic account of knowledge. On the other hand, an assumption that an "architecture that supports knowledge" is an architecture that supports internal encoded symbolic representations is an ultimately incoherent, untenable, and unworkable assumption — an assumption that will *never* yield genuinely knowledgeable systems. To this point, then, we are in full agreement.

The potential problem here is that an analysis of knowledge that is fundamentally observer-dependent avoids the incoherence of construing *internal* system representation as encodings, but it threatens to *de-naturalize* all knowledge. As a result, rather than being part of the natural world, knowledge would require — ontologically require — an observer. At worst, such an analysis could lead to a full observer idealism.

Clancey's statements lead to the brink of this problem. If we take into account his earlier point that analysis should attend to the problem of knowledge for a robot, and if we interpret "knowledge" in the above statements to mean "AI type knowledge level descriptions," then he, specifically, does not cross the line. The *general* point remains, however, that too much "observer talk" — "observer talk" that makes unrestrained ontological commitments to observer dependencies — has devastating consequences for theories that purport to explain the nature of representation. We have already encountered a full observer idealism in the work of Maturana & Varela, and we later see a similar trend in the discussion of Winograd & Flores (1986). Clancey's comments highlight this danger.

Resolution: A Category Error and Its Avoidance. There is, however, a resolution to this interpretive problem, that does not necessarily commit to idealism, and that provides its own interesting perspective on the issues. Clancey's situated cognition and interactivism are in full agreement that there is a fundamental difference of kind between external representations (such as pictures, maps, blueprints, and so on) and whatever it is that constitutes internal mental representation. External representation and internal representation — the nature of intentionality — cannot be the same kind of phenomena. Simply, external representations require interpreters — the people using them as representations. This is not a problem in principle for pictures, maps, and so on. But internal representation cannot involve interpreters.

Internal representations cannot require interpreters, because mental representation is an aspect of a person, an aspect of *the nature **of** an interpreter*. For internal representation to require interpretation is to commit to an infinite regress of interpretive homunculi each interpreting the encodings of the preceding homunculus into still new encodings. The assumption that the mental representation problem can be solved by a model of internal versions of external representations commits directly to encodingism. Models of mental phenomena in terms of external sorts of representation, then, such as symbols and data structures, are not only factually wrong, but are infected with the logical incoherence, and consequent infinite regress of homunculi (among other corollaries), of encodingism. This point reinforces the convergence between the criticisms from within situated cognition and those of interactivism.

Since *external* representations require interpreters, and foundational *internal* representationality *cannot* require interpreters — on pain of infinite regress, among other consequences — to attempt to model internal representationality in terms of such external kinds of representations is to commit a category error (Clancey, 1993). This point that standard approaches involve a category error is still another perspective on — another member of — the group of corollaries of the incoherence argument.

A Machine Language Rejoinder. One response to this point might be to argue that machine language is not interpreted, but, rather, simply enacted. Thus, so the point continues, the regress of interpretation bottoms out innocently, and the physical symbol hypothesis lives very nicely. There are several possible quibbles with this argument, but the basic problem with the response is that machine language per se contains *no representations*. A machine language program is "simply" a machine specification — in terms of multitudinous switching relationships, control flow commands — that the computer then simulates. So, the problem of how higher level "representations" are made good in terms of control-flow machine language is not even addressed. Machine language captures at best a (limited) kind of "knowing how" and leaves all issues of "knowing that" untouched.

The implicit claim is that such "knowing that" is somehow emergent out of lower level machine language "knowing how." We are clearly in strong sympathy with this point in its *general* form, but the standard view gives no account of how such an emergence is supposed to occur, and the assumed nature of "knowing that" as encodings *cannot* be

directly emergent out of "knowing how," so the assumptions concerning the two sides of the emergence relationship are logically incompatible — the overall position is incoherent. All "knowing that" claims for higher level languages are still redeemed, at best, in terms of correspondences — that is, in terms of encodings — not in terms of "knowing how." That is, they are "redeemed" by the category error of assuming that external representations requiring interpreters are the fundamental form of internal representationality (Bickhard, 1992c, 1993a).

How to Talk. Given the fundamental difference in kind between internal and external representation, together with the standard presupposition that there is in fact no such fundamental difference (and, thus, no standard distinction in the language used to talk about the two cases — e.g., Vera & Simon, 1993) there is a choice about focus and about language usage. Interactivism focuses on the nature of internal mental representation, and continues to use representational language to discuss it, though with multiple qualifiers, caveats, and logical arguments concerning the necessary ontological distinction between internal and external representational phenomena. Clancey and other situated cognitionists have focused on external representations and their necessary ontological difference from mental intentionality, and have *eschewed* the usage of representational language for such internal, mental, phenomena — the mental phenomena *cannot* be the same sort of thing as the external notations and representations (Slezak, 1992, 1994). In this view, (external) representation is a kind of use, an interpretive use, that people make of various things in their environments, and several kinds of things that people create in their environments *in order* to be able to make such representational uses of them. But such use and creation of external representation must not be confused with what actually might be going on in the mind or brain — they cannot be the same sort of phenomena.

On this understanding, confirmed by Clancey (1992d, 1993), there is no fundamental disagreement between interactivism and Clancey's situated cognition. There is, however, a difference of focus and vocabulary usage — one that can be potentially confusing. Nevertheless, this perspective of situated cognition provides still another view on both the errors of encodingism, and the seductive powers of encodingism.

A General Note on Situated Cognition

Situated cognition as a general development within Artificial Intelligence proceeds from, among other sources, valid and deep insights

concerning the differences between external and internal representational phenomena. We emphasize, however, that the correct characterization of external representation as ontologically requiring an interpreter, an observer, can easily be (mis)understood as, or even logically commit to, an ontological necessity for an observer for *all* representational phenomena, internal as well as external. In this form, the commitment is to an observer idealism, which, as pointed out in discussion above, is not only itself in error, but is itself a version of the encodingism error. To presuppose that internal representation requires an observer, as does external representation, is a violation of the ontological distinction between external and internal representational phenomena. It is to assume that *all* representational phenomena require interpretation, and that assumption is an encodingist assumption.

Overemphasis on the necessity of an observer or designer or constructor and consequent implicit or explicit assimilation of internal representation to this insight concerning external representation, risks committing to idealism. It is our judgment that some situated cognitionists have, wittingly or unwittingly, engaged in the inconsistency of straddling this issue, and others have, wittingly or unwittingly, crossed over it into a full idealism. We note once again that such an idealism is itself just another version of encodingist presuppositions.

Rodney Brooks: Anti-Representationalist Robotics

Brooks proposes a radical shift in approaches to the construction of artificial intelligence (1990, 1991a, 1991b, 1991c, 1991d). He suggests that the problem of artificial intelligence has been broken down into the wrong subproblems, and in such a manner that it has obscured the basic issues and approaches to solution. He proposes that intelligent systems be constructed *incrementally*, instead of *componentially*, starting with simple intelligences and working toward more complex instances, *with each step constituting a full intelligent creature* "in the real world with real sensing and real action." (1991a, 140)

In following this approach, Brooks and colleagues have arrived at an unexpected conclusion and a radical hypothesis:

> Conclusion: "When we examine very simple level intelligence we find that explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model."

> Hypothesis: "Representation is the wrong unit of abstraction in building the bulkiest parts of intelligent systems." (1991a, 140)

In fact, Brooks suggests that "Representation has been the central issue in Artificial Intelligence work over the last 15 years only because it has provided an interface between otherwise isolated modules and conference papers." (1991a, p. 2)

Subsumption and Evolution. In the construction of intelligent systems, Brooks advocates a layered approach in which lower layers handle simpler behaviors and higher layers handle more complex behaviors, generally through influencing the activity of the lower layers — the higher layers *subsume* the lower layers. By getting the lower layers to work first, debugging and correcting the next layer up is enormously simplified. Brooks is proposing to model his engineering approach on evolution (1991a, p. 141), both in the construction of individual intelligent creatures, and in the design and construction of new intelligent creatures. (He notes that the problems that evolution took the longest to solve — those of basic real world interaction — are precisely the ones that standard Artificial Intelligence deliberately ignores.) The successive layering that results recapitulates both evolution *and* physiological design and maturation (Bickhard, 1992b).

With the extremely important caveat that "representation" in Brooks' discussion be understood as "symbolically encoded representation," we are in enthusiastic agreement with Brooks. In fact, his analysis of "representation" emerging only because it is needed for the interface between otherwise isolated modules is virtually identical with the interactivist analysis of the emergence of subsidiary encodings and associated processing systems out of underlying interactive systems (Bickhard & Richie, 1983). There is a significant difference in emphasis, however, in that the interactivist analysis is of how and why such differentiated modules and encoding interfaces might evolve and be functional *in the context of and in the service of an already functioning non-encoding interactive system*. That is, might evolve and emerge in the service of exactly the sort of intelligent systems that Brooks proposes to create.

Robotics. Robotics has at times been seen as "simply" a subdivision of Artificial Intelligence. Brooks' proposals in effect reverse that. He suggests that, by virtue of having to interact with the real world, robotics in fact encounters the fundamental problems of intelligence, and, conversely, that the isolated module approach standard in Artificial

Intelligence cannot do so. This is also exactly the point argued in Bickhard (1982). It is clear that robots are *necessarily* interactive, while standard Artificial Intelligence is *deliberately not*.

Robotic Representations. There is, of course, one major difference between the interactivist approach and the position taken by Brooks: by interactivist criteria, at least some of Brooks' intelligent creatures already involve primitive representations. He acknowledges that there might be some sense in which representations are involved "implicitly" (1991a, p. 149), but argues that what goes on in such creatures is simply too different from "traditional AI representations" (1991a, p. 149) and "standard representations" (1991a, p. 149) to be considered as representations. This point is quite correct in its premises, but by thereby dismissing the possibility that intelligent creatures *do* involve representations in some real, but *non*-standard sense — an interactivist sense — Brooks inhibits the exploration of that aspect of his project.[10] Interactivism, in fact, argues that representation first emerges in *exactly* the "implicit" sense that Brooks reluctantly and indirectly acknowledges, with all instances and forms of explicit representation — "standard representation" — emerging from, and remaining subsidiary to and dependent upon, such interactive "implicit" representation.[11]

Brooks (1991d) provides a first step in this interactivist direction when he proposes an "inverse" perspective on the function of sensors. In the standard perspective, in which the focus is on inferring the correct world state on the basis of sensor readings, the designer focuses on "[a] particular world state and then analyz[es] all possible sensor readings that

---

[10] There are technical niceties involved here concerning the exact boundary requirements for something to be a representation. The basic point, however, is that the domain of 1) functional indications of possible interactions, 2) functional indications of subsequent internal outcomes, 3) selections of interactions on the basis of such indications, 4) influence or control over subsequent process on the basis of the success or failure of such indication of internal outcome, and so on, is the domain of the emergence of interactive representation, and that such functional organizations are present in, and are easily compatible with, Brooks' robots (e.g., Mataric, 1991). In rejecting and disclaiming representation, Brooks is quite correct in standard senses, but is overlooking such interactive possibilities.

[11] Notice that Brooks' creatures constitute a concrete counterexample to the "all or none" presupposition about mental phenomena. They illustrate real representation in a form that is decidedly not at the level of human intelligence. This separation may be another reason why Brooks looks so radical to standard AI.

A recognition of primitive representation, of course, is inherent in any evolutionary perspective that recognizes any kinds of mental properties in simpler organisms. Brooks shares this with a number of other researchers that have a biological orientation, such as Krall (1992) or many people within the dynamic systems approaches.

could be generated" (p. 438). In Brooks' inverse approach, the designer proceeds to consider "given a sensor reading, ... which possible worlds could have given rise to that reading" (pp. 438-439).

The relationship between a sensor reading and the "possible worlds [that] could give rise to that reading" is *precisely* that of interactivism's implicit definition (Bickhard, 1980b, p. 23 — see also Bickhard, 1992c, 1993a, in preparation-c; Bickhard & Campbell, 1992, in preparation; Campbell & Bickhard, 1992a). Four additional steps are required to arrive at the interactive model of representation:

1) Recognition that this relationship is *only* implicit from the perspective of the system itself;[12]
2) Generalization of such implicit definition from strictly passive sensory input receptors to interactive (sub)systems;
3) Recognition that this implicit definitional relationship does not provide an emergence of representational content for the system itself; and
4) The addition of the implicit predication of further interactive properties as providing that emergence of content.

Agre and Chapman

The work of Agre and Chapman, despite being motivated by a different set of initial considerations than interactivism, manifests a remarkable convergence with interactivist notions of representation and activity (Agre, 1988; Agre & Chapman, 1987; Chapman, 1987; Chapman & Agre, 1986). The basic orientation that they represent is that of situated cognition — they share programmatic frameworks with others such as Brooks, Clancey, and Smith.

Heideggerian Parallels. One major motivation for Chapman and Agre's development has been the Heideggerian phenomenological realization of the ongoing dynamic of everyday life — perception and action function continuously and simultaneously, and decisions occur as aspects of the basic processes of perception and action in dealing with a changing environment.

Planning is Inadequate. In particular, there is *not*, except in unusual circumstances, a separation of stages of Input, Planning, Decision Making, and Action. Complementary to this realization is the powerful

[12] This recognition is already "implicit" in Brooks' discussion: He is writing from within the perspective of analysis by a designer, not from the system's perspective.

negative result that Planning, in the above detached-from-the-environment computational sense, is computationally intractable (Chapman, 1987). In other words, the activity of life proceeds in continuous interactions — with perception, decisioning, and action as aspects rather than parts or stages — and, furthermore, detached Planning of the sort typical of AI research (Fikes & Nilsson, 1971; Hendler, Tate, & Drummond, 1990; Miller, Galanter, & Pribram, 1960; Sacerdoti, 1977; Sussman, 1975; Warren, 1974; Wilkins, 1988) is fundamentally impossible anyway, since it presumes a complete encoded world model upon which computations that determine action operate. Classic Planning theory is both epistemologically (because encodings cannot do the required representational work), and heuristically (because of the intractability of the problem of selecting courses of action by computing over an encoded model), inadequate as a theory of action (Agre 1988).

Generalization. Implementing such insights, however, encounters a foundational problem: standard approaches to representation. In particular, the dominance of the encoded name as the paradigmatic form of representation, together with the dominance of the approach of detached computation on totally internal representations, has produced an AI culture in which each "object" in the "world" is encoded by its own unique name. With simple enough "worlds," and so long as the programmer is willing and able to artificially provide all such names, this appears tractable in itself. It creates still another problem, however: the classic problem of encodingism — the problem of generalization.

Specifically, any procedure learned with respect to any particular objects must be explicitly generalized to other appropriate objects. It would be intractable indeed if every simple routine needed to be relearned for each new set of objects. The standard approach to this problem is to render the procedures in terms of variables in place of names, and then allow generalization to new objects in terms of replacing the variables by new names — or some equivalent. This requires the abstraction of the form of a procedure away from the particular objects with respect to which it has been initially constructed, generalizing to a "pure" form in terms of variables, and then reparticularizing with new objects. These abstracted forms constitute patterns that might or might not be instantiated. Such pattern matches must be searched for and, when found, the appropriate procedures executed.

Unfortunately, the problem of such abstraction is not in general solved, and the problem of computationally detached pattern matching is

itself a version of computationally detached Planning or Problem Solving, and it too becomes quickly intractable. All such detached computations are, in addition, frightfully expensive to set up as well as to compute.

Deicticness. At this point, a second influence of the phenomenology of Heidegger is felt. Instead of encoding particular objects with individuated names, the notion is found of encoding *deicticly*. That is, to represent the world *indexically* — with respect to the agent's body — and *functionally* — with respect to the agent's purposes. The ultimate point, after all, "is not to express states of affairs, but to maintain causal relationships with them." (Agre, 1988, 190-191; Chapman, 1991) Representation is foundationally functional, and the functions are in the service of *interaction* — interaction of *this particular agent*.

Deictic representation has numerous advantages. Paramount for the above considerations is the sense in which deictic representations provide an intrinsic abstraction and generalization. In particular, a procedure in terms of "to the front[indexicality] is a potentiality for <eating, grasping, becoming-hurt, and so on>[functionality]" is indifferent to the differences among the innumerable potential instantiations of that potentiality, and, therefore, *will intrinsically generalize over all of them*. Such transfer is essential to any realistic interactive system.

Interactivism and Deicticness. Here we find several strong convergences with interactivism. First, interactive representation is constituted as open differentiations by the agent of the agent's environment. These differentiations have representational content, insofar as they do, in terms of the potentialities for further interaction that they indicate. That is, they have indexicality with functionality.

Interactive differentiations and interactive representations are *intrinsically* deictic, indexical, and functional. Differentiations cannot be other than indexical — relative to the agent — and content cannot be other than functional — relative to the agent's potential activity. Agre and Chapman's notions of situated cognition have here converged with fundamental properties of interactive representation.

Ongoing Interaction. Furthermore, the recognition of the ongoingness of living interaction, and the fundamental sense in which representation is in the service of such interaction (activity), are both common recognitions and common motivations between the two approaches. Still further, Agre and Chapman propose a sense in which

abstract reasoning might be emergent from concrete activity (Chapman & Agre, 1986) that seems very similar to the interactivist proposal for the construction of derivative encodings and related reasoning procedures on the foundation of and in the context of a general interactive system (Bickhard & Richie, 1983). In such respects, there are striking convergences between interactivism and the positions of Agre and Chapman.

In recognition of the ongoingness of real living activity, Chapman and Agre (1986) argue against "detached computation" in the sense of computations lifted out of this domain of real activity. With regard to such *critical* stances there are also convergences with interactivism. In particular, encodingism sunders representation from interaction, and then must try to *reconnect* encodings, with interpreters that bridge from data structures to action systems.

The action independence of encodings is another aspect of their general context independence, and, in being context independent, they cannot make use of nor rely on resources in those contexts. Thus, they must attempt to anticipate everything, to have complete internal world models. Conversely, they have an impossible task with real time, *always changing*, environments. This chain from encodings to detached computational models is not a logically valid one — it is not logically forced — but, rather, it is a chain of motivations, or, better, a chain of implicitly guided focuses of attention with critical hidden presumptions. The field, nevertheless, seems to have followed these considerations fairly universally, with the "detached computation" results that Agre and Chapman (and Brooks: see above) are now reacting against. Thus we find still another convergence between Chapman and Agre and interactivism.

Differences. There are differences, however. We are in strong agreement with several of the themes of the situated cognition position — such as embodiedness, situatedness, and an emphasis on dynamics — but the details ultimately do not succeed in avoiding the problems of encodingism.

Deictic Encodings. Most importantly, the deictic representations that are proposed in this approach are, nevertheless (and in spite of the fundamental insights inherent in and advantages of such deicticness), *encodings*. "To the front" is deictic, but it is not an emergent representational content generated by the system itself; it is a designer provided encoding. Perhaps the simplest way to see this is to note that

there is no account of representational *content* in Chapman and Agre's situated cognition. There are signals that affect activity, but no *emergence of representation for the system itself.* From the interactivist perspective, there could *not* be representational content because of the absence of *goals*. Activity systems — Agre's dependency networks (Agre, 1988) — may accomplish goals from the observer perspective, but that is implicit in the observed relationships between system activities and the environments — it is not a condition functional in the system itself. A dependency network can achieve environmental conditions, but it cannot be error detecting or correcting except insofar as the space of possible errors and appropriate responses are all explicitly anticipated and encoded into the dependency network. Even in such a case, there are still no "errors" — failures to achieve goals — from the system perspective, but only various inputs and dependencies.

The encodingism of situated cognition is also implicit in the choice of deictic *entities* as the "references" of the encodings (Agre, 1988, p. 191). *Entities* in this sense are indexically and functionally specified objects, but a single object might be referenced as more than one entity so long as it fit into more than one functional category. This move to indexical, functional entities is a vast improvement over objectively named objects, but it is not a move all the way to interactive properties. From the interactivist perspective, *all* representation must be constructed from basic differentiations and predications of *interactive* properties, with relative epistemological objectivity emerging as various forms and degrees of *invariances* of organizations of interactive properties relative to other interactions. Piagetian object permanence is a paradigm of such invariance (Piaget, 1970b, 1971, 1985).

Absence of Goal-Directed Interaction. Without moving to such a level of interactive properties, no model of representation can connect with the basic interactive system organization, and, therefore, cannot account for the emergence of more invariant, stable, forms of representation. Without moving to such a level of predications of interactive properties, a model of representation is forced to define its representations in terms of what is being represented, rather than in terms of how those representations *emerge* — but representations defined in terms of what they represent *are* encodings. It should be noted that this move to a deictic, indexical, functional level, *but not beyond*, and the consequent implicit commitment to encodingism, is characteristic of

Heidegger in *Being and Time* (1962). Heidegger seems likely to be its source in Agre and Chapman's positions as well.

This absence of goals in this approach is due in part to an esthetic heuristic of machinery parsimony — make due with as little machinery as possible — and motivated still further by the sense in which " 'state' in AI research has almost always meant 'world model' " (Agre, 1988, 247), which returns directly to the problems of detached computation. This model, then, uses stored states only when forced to. Since much activity can be accomplished without stored states, it is heuristically eschewed, and, without stored states, there can be no goals to be sought — therefore no interactive representational content, and, therefore, the deictic representations are encodings, with contents (and goals) known only to the designer or observer.

Interactivism is motivated more biologically (Bickhard, 1973) and, in living systems, the existence of survival-directed, thus goal-directed, organization is intrinsic in all but the most marginal forms of life — perhaps viruses. Questions of parsimony of mechanism take on a different force when some sort of mechanism is *already known to be available*. When a type of mechanism — such as goal-directedness — is already known-to-be-present, then deviations that *eschew* such mechanisms become violations of parsimony that require their own additional justification. The esthetic heuristic of mechanism parsimony, then, cuts very differently in interactivism than in Agre and Chapman's situated cognition (Agre, 1988; Chapman, 1991).

From the interactivist perspective, *extremely well habituated* interactive procedures may *partially* approximate the reactively and ongoingly run-off interactions of Agre's dependency networks — with minimal explicit goal-directed process. Of course, from Agre's perspective, one of the main points is precisely that much of daily life *is* extremely well habituated. But if the interactive model of representation and representational content is correct, then goal-directed organization of a system is essential to the emergence of representation.

The absence of goal-directedness in Agre's model would seem to be an asymptotically unreachable limiting case even given thorough habituation — at least for any complicated organisms with nervous systems, and certainly for human beings — for other reasons in addition to the emergence of representational content. For one, since various levels of the nervous system are themselves organized as layers of

servomechanisms, the complete elimination of goal-directedness, no matter how deep the habituation, seems physiologically impossible.

Furthermore, and more deeply, the unavoidable unreliability and ambiguity of sensory signals that are correctly held against detached computational systems (Agre, 1988) would also seem to preclude *in principle* the total elimination of goal-directedness, even in relatively simple systems. This is so because the power of goal-directed differentiations of the environment is needed to overcome that sensory unreliability and ambiguity — for example, visual interactions, such as eye or head or body motions, for the sake of parallax effects providing depth information that is unavailable in simple retinal images (Bickhard & Richie, 1983).

Similarly, the unreliability of *actions* requires goal-directed corrections. Action virtually always proceeds by way of progressive approximation to the desired goal, and no actions are precise enough and predictable enough in their outcomes to avoid such a process. There are always conditions of fatigue, immobilized limbs (e.g., it's carrying something), unexpected blocks (e.g., the bicycle is in the way), and so on that preclude non-corrected activity (Maes, 1990b). Most fundamentally, without goals, there can be no error for the system, and, therefore, no representational content for the system.

Deictic Abstraction. Another contrast with this particular model of Agre's has to do with its characterization of the abstractness of deictic representation as "passive abstraction" (Agre, 1988). This is not incorrect, but it is unfortunate in that it seems to imply that abstraction and consequent generalization is still something that "gets done," it's just that it "gets done" passively. From the interactive perspective, however, the abstract character of interactive differentiations is *intrinsic*, not passive. The abstractness of differentiations does not "happen," it is an intrinsic aspect of the nature of differentiations. That is, anything that is *not explicitly* differentiated *is implicitly* abstracted and generalized over. From the perspective of deictic *encodings*, however, a notion of passive abstraction has a little more appeal.

Neither the machinery parsimony nor the term "passive abstraction" involves any deep assumptions or commitments of Chapman and Agre. We mention them because their *presuppositions* do involve and reveal more serious potential differences with interactivism. Minor as they are, they are positions and usages that could not be coherently taken from within interactivism.

Interactivism and the situated cognition of Agre and Chapman share recognitions of the ongoingness of every day life, of the fact and importance of deictic representation — and of some of the deep relationships between the two. For both, taking seriously how action and interaction is actually accomplished in the world yielded powerful constraints on theory not normally taken into account in Artificial Intelligence or Cognitive Science — constraints that invalidate much of contemporary and dominant approaches. A significant portion of this convergence arises because interactivism is necessarily and intrinsically situated. Only a situated interactive system can instantiate interactive representation, and that representation is intrinsically system anchored and functional in nature: deictic and indexical. Interactivism suggests, however, the necessity of internal state goal-directedness to be able to account for the emergence of representational content, and, thus, to be able to avoid the aporias of encodingism.

Benny Shanon

Context Dependence. Benny Shanon presents a powerful argument against the viability of standard models of representation (1987, 1988). The fulcrum of his argument is the context dependence of meaning. Such context dependence has certainly been noticed elsewhere, but what Shanon points out is that, unless relevant contexts were somehow limited to some small set of categories, context dependence makes *impossible* any rendering of semantics in terms of fixed, finite representational elements — in terms of anything like a language of thought (Fodor, 1975). Contextual variation in meaning is, in principle, unbounded — even with regard to single words — and this is beyond the explanatory power of any elemental or atomistic approach. The logic is simple: finite sets are not capable of capturing unbounded variations.

Shanon's conclusion is that representations are the *product* of cognitive work, not the *basis* for it. They are generated in-context as appropriate *to* that context. In this sense, representations are to be *explained* by a model of cognition, instead of serving as the presupposed *ground* for such explanations.

Two-Stage Understanding. Shanon considers and rebuts several variants of a possible rejoinder to his argument from contextual variation. The basic rejoinder is the standard two-stage model of understanding. It claims that meaning is fundamentally understood, and first understood, in terms of literal, fixed, elemental — context independent —

representational meaning. This, supposedly, is then followed by some sort of process of modifying that initial, literal understanding that takes context into account. Shanon looks at two versions of such models: one for metaphor, and one for indirect speech acts. He points out that the two-stage notion is supported neither by processing time studies, nor by developmental data — in neither case does literal understanding come first.

Furthermore, both literal *and* non-literal usages show context dependencies. For example, consider the "literal meaning" sentence "You are going to lose all of your money." where context determines whether the money is in your pocket or your investments. Even worse, the *very distinction* between literal and non-literal is itself context dependent, and even obscure. Consider, for example, "All men are animals" spoken by a biologist or by an angry feminist. Or consider "The policeman held up his hand and stopped the car." compared to "Superman held up his hand and stopped the car." Which one involves the literal meanings, and which one the non-literal?

Shanon concludes that two-stage models fail both empirically and conceptually, returning the discussion to the unbounded context dependent variations that cannot be captured in finite sets of atomic meanings. He suggests that context-dependent pragmatics is primary, rather than secondary, and is the matrix out of which semantics emerges.

Transcending Chomsky. We would like to point out an interesting and ironic parallel between Shanon's argument against representationalism and Chomsky's argument against associationism. In Chomsky's argument, associations are computationally inadequate to the unboundedness of the number of possible sentences, while rules are computationally adequate. In Shanon's argument, atomic representations are computationally, combinatorically, inadequate to the unboundedness of the number of possible contextual variations of meaning, while context embedded actions are computationally adequate.

Convergences. With two caveats, we enthusiastically endorse and support these remarkably convergent points of Shanon's. These convergences include a ubiquitous context dependence and a recognition of the primacy of pragmatic process over atomic atemporal encodings.

Context Dependence. Context dependence is an intrinsic characteristic of the interactive model. Interactivism, in fact, manifests *two* senses in which utterances are intrinsically context dependent, and, thus, subject to contextual variation. First, utterances are contextually

interpreted *operations* on their social reality contexts, and, therefore, their *results* depend as much on that context as they do on the operations engaged in. Second, those social realities are themselves constituted out of the participants' representations, and those representations are contextually open, deictic, interactive differentiations. There is no restriction to an encodingist set of atomic representations in this model, and the context dependencies involved are *intrinsic* to the nature of what utterances and representations *are*. Among other consequences, that means that there is no need for postulating an ad hoc second stage of context dependency on top of a first stage of context independent decodings.

Pragmatics is Primary. Furthermore, the interactivist models of both representation and language support Shanon's sense that pragmatics is primary, not secondary. Both interactive representation and utterances are, and emerge out of, pragmatic action and interaction. It should be noted, however, that this involves the notion of pragmatics in the primary sense of "pragmatic" or "pragmatism": the standard *linguistic* distinctions between syntax, semantics, and pragmatics are already logically committed to an encodingism (Bickhard, 1980b, 1987).

Still further, interactivism offers a solution to the problem of the genuine *emergence*, the *creation*, of representation — in evolution, development, learning, cognition, and perception. Interactive representation emerges out of the functional relationships in the organization of a goal-directed interactive system, and *secondary encodings* can be constructed in the service of relatively (and partially) ***de***contextualized, generalized, heuristics and abilities — such as, for example, explicit inferences (Bickhard & Richie, 1983). Both representat***ion*** *and* representat***ions***, then, are the products of cognitive work.

Even Shanon's argument here — that representations must be the product of cognitive work, and not the basis for cognition — parallels the interactivist argument that encodingism presupposes the fundamental phenomena of representation — hence, a programmatic encodingism is explanatorily circular. In fact, it is by now a commonplace observation that neither Artificial Intelligence nor Cognitive Science can account for the representational content of their encodings. Hope for a solution to this aporia seems to have contributed to the excitement over connectionism. We show below, however, that connectionism does not offer a solution.

Caveats. We do have a couple of caveats to our agreement with Shanon's positions. They are, however, quite minor.

Utterances. The first caveat to our endorsement stems from the fact that most of Shanon's examples are of utterances. This is standard practice when discussing meanings and representation, and, in making his case against standard representationalism, Shanon is quite justified in taking utterances in their standard guise. That standard interpretation, however, involves, among other things, the assumption that utterances are *themselves* representational — recodings of cognitive encodings. Interactivism, on the other hand, models utterances as *operations on* (social realities which are constituted as) representations. In this view, utterances are no more representational than functions on the integers are odd or even or prime or non-prime. There is most certainly a representational *aspect* of utterances — representations are (constitutive of) what utterances operate on — but to render the utterances themselves as representational is already to commit to their being defined as representations in terms of what they represent — that is, to commit to their being encodings.

Connectionism. The second caveat is simply that we do not endorse Shanon's cautious suggestion that perhaps connectionism offers a way out of these difficulties. Connectionism offers definite strengths compared to standard symbolic encodings — as well as weaknesses — but does not escape the common assumption of encodingism (see below). See, however, Shanon's more recent and more negative assessment of connectionism (Shanon, 1992).

Representational-Computational View of Mind. More recently, Shanon (1993) has integrated these critiques with a range of others to produce a massive exploration of the failings of what he calls the representational-computational view of mind. He addresses, for example, the problem of the interpretation of encodings: If interpretation is into something that is not a structure of encodings, then encodingism is not the basic form of representation, but if interpretation is into other encodings, then these too must be interpreted, and we have an infinite regress of interpreters and interpretations. He points out the impossibility of the origins of encodings — the problematic that yields Fodor's claims of innateness (Fodor, 1981b) — and the unbridgeable epistemic gap between encodings and the world — the problematic that yields Fodor's solipsism (Fodor, 1981a).

Shanon is also concerned about the atemporality of encoding representations. However much it may be the case that encodings are set up and manipulated in time, and however much it may be the case that we may care how much time such processing may require, it remains true that nothing about what is supposed to make encodings representational — nothing about the correspondences that are supposed to constitute the encodings — is inherently temporal. The presumed encoding relationships would exist, if they exist at all, whether or not there were any temporal processing going on. This is in stark contrast to the fundamental temporality of human psychology, and in flat contradiction to the intrinsic temporal modality of interactive representation.

Shanon discusses why representationalism is such a powerful framework, in spite of its many failings, and addresses many additional issues concerning representational-computationalism. He also develops a set of constraints on a more viable model of representational phenomena. The focus of these constraints is on cognition as a form of action — and action and interaction by an embodied agent embedded in the world. Clearly there are core convergences here with the interactive model. It is not as clear, however, whether those convergences extend to such properties as implicit definition, implicit predication, and the system detectable truth values to which they give rise.

Motivation. Shanon addresses one topic that we would like to explore a bit further from the interactive perspective: motivation. If motivations are encoded in standard manners — as goal descriptions, perhaps — or as motivational tags on other representations, then, in standard manner, they need interpretation. In this case, however, the interpretations are not only semantic but also "motivational" — whatever motivation is, it is constituted in the interpretation, not the encoding per se (just as with semantics). The encodings are themselves inert, and have no independent action or energizing properties. Motivation has to be somehow tacked-on to atemporal representational structures and relationships.

On the other hand, if motivation is construed as some sort of energizing of representational structures, some sort of motion or change or putting-into-action of inert representations, then we have introduced a process that cannot be captured within the computations-on-representational-items framework. Any such motivational processes cannot simply be more of the same of manipulations of representations, because motivational processes must eventuate in, among other things,

action. Action, and action initiation and selection — i.e., motivation — are outside the scope of standard frameworks. Shanon points out that the postulation of a third component, a motivational process — in addition to representations and computations on representations — introduces the possibility that the original pair, representations and computations, is poorly defined and unnecessary. Motivational processes may *include* what are standardly modeled as computational processes, and representational phenomena may be *implicit* in such motivational-computational processes. Distinct categories of representational elements and dedicated computational processes may be a false subdivision of the study of mind. If representation is an emergent of action, and action selection is inherent in the organization of an interactive system, then the boxology of representations, computations on representations, and motivational energizers — pushers and pullers — is misguided.

Action Selection. Any living system is necessarily a system in motion. Living systems are open systems, and open systems are *processes*. If an open system stops, it no longer exists. Living systems, then, are always doing something; doing nothing is death. The question of motivation, therefore, is not one of "What makes the system do something rather than nothing?" — it *must* be doing something by virtue of being alive — the question of motivation must rather be one of "What makes the system select, or what constitutes the system selecting, one thing to do rather than another?"

Interactive representation is a functional emergent of interactive open systems, and *function* — in this model — is itself an emergent of open systems, including in particular living systems (Bickhard, 1993a; Bickhard & D. Campbell, in preparation). More specifically, interactive *representation* is a property, an aspect, of the organization in an interactive system for *selecting* next courses of interaction. One aspect of the organization of an interactive system, then, will involve indications of potential interactions — a representational aspect — and another aspect will involve selections of interaction based upon, among other things, such indications — a motivational aspect. In this model, representation and motivation are different aspects of the same underlying ontology of interactive dynamic systems, just as a circle and a rectangle are aspects of one underlying cylinder (Bickhard, 1980a, 1980b; Campbell & Bickhard, 1986). Much more, of course, needs to be developed here, including such phenomena as "drives," pain and pleasure, emotions, the emergence of higher order motivations such as curiosity and esthetics, and so on

(Bickhard, 1980a, in preparation-b; Campbell & Bickhard, 1986). The critical point to be made in this context, however, is that, in the interactive model, motivation and representation are not dirempted from each other, but, instead, are united as different aspects of the same underlying ontology. This stands in fundamental contrast to standard representationalism — and in fundamental agreement with Shanon (1993).

Pragmatism

Interactivism shares many of its basic points, both critical and constructive, with pragmatism (Murphy, 1990; Thayer, 1973). On the critical side, pragmatism presents a decisive rejection of the correspondence approaches to *truth* and *meaning*. This passive notion of knowledge and its origins is dubbed a *spectator* epistemology (J. E. Smith, 1987). Peirce proposed levels of analysis of *experience* and *meaning* that supersede sense data — for example, we do not *perceive* the Doppler shift of a train whistle as the train goes by, but we do *experience* it (J. E. Smith, 1987). Experience and meaning became primary domains of inquiry for later pragmatists.

On the constructive side, experience and meaning were construed in pragmatic terms, as involving anticipations of consequences. Revisions of meanings, in turn, occurred in response to failures of those anticipations (Houser & Kloesel, 1992; Thayer, 1973). Here is a clear anticipation of interactivism's focus on interactive consequences, and the consequent variation and selection constructivism. Furthermore, again in agreement with interactivism, pragmatism involves a commitment to process as a central aspect of its metaphysics (Bourgeois & Rosenthal, 1983; Houser & Kloesel, 1992; Murphy, 1990; Thayer, 1973; Rosenthal, 1983, 1987, 1990, 1992; Rosenthal & Bourgeois, 1980).

Peirce, however, retained an essentially correspondence notion of representation *per se* in his semiotics (Hoopes, 1991; Houser & Kloesel, 1992; J. E. Smith 1987; Thayer, 1973). Experience and meaning became the more important domains of analysis, but encodingism was not rejected. Some later pragmatists have identified the very notion of representation with correspondence, and have rejected it, leaving *only* experience and meaning as relevant domains (J. E. Smith 1987; Rorty, 1979, 1982, 1987, 1989, 1991a, 1991b).

Idealism. A total rejection of representation, however, is simply a move to idealism. This is not a welcome consequence: Rorty, for

example, claims that his position is not an idealism, that nothing in the rejection of representation precludes the fact the people have to get along in the world. Getting along in the world, however, involves the possibility of *failing* to get along in the world, and of learning how not to fail. Learning how not to fail in interaction *is* the construction of representation. Rorty provides no way to model such learning-to-avoid-error, no way to understand an epistemic encounter with error, and no way to model what is constructed in the process of such learning. He equates even taking *questions* of representation seriously as being committed to the rejected correspondence models (Rorty, 1979, 1982, 1987, 1989, 1991a, 1991b). He doesn't want idealism, in other words, but he provides no way to understand our world within the confines of what remains from his totalizing rejections.

Peirce, on the other hand, is a pragmatist who did not reject representation, but the result of his efforts was yet another encodingist model of representation — a very sophisticated version to be sure, but encodingist nevertheless. Interactivism claims both to model representation and to do so without epistemic correspondence.

Ken Ford. Pragmatist intuitions have not played major roles in Artificial Intelligence or Cognitive Science. Their strongest impact has been on evolutionary models of learning, not on models of representation per se. One exception, however, is the proposal of Ford and his collaborators (Ford, 1989; Ford & Adams-Webber, 1992; Ford & Agnew, 1992; Ford, Agnew, & Adams-Webber, in press). Following Kelly (1955), a cognitive personality psychologist who was influenced by pragmatism, they point out that the fundamental function of representation is anticipation, and that the only way available for modifying representations is in terms of their failures of predictive utility (Ford & Adams-Webber, 1992). Here, again, we find the emphasis on consequence, and on an error driven constructivism — both strongly convergent with the interactive position.

Kelly, however, along with pragmatism in general, accepts a basic epistemic ground of uninterpreted contact with reality — an "uninterpreted given" of sense data, quale, the flow of reality, events, and so on, with the exact terminology varying from pragmatist to pragmatist — on top of which, and with respect to which, processes of judging, interpreting, construing, anticipating, and so on are proposed to occur (Houser & Kloesel, 1992; Husain, 1983; J. E. Smith 1987; Thayer, 1973). In a sense, all the "action" is in the levels of anticipatory interpretation,

and this is what is most focused upon. But it is only in terms of the uninterpreted "givens" that the anticipations and failures of anticipations of the realm of experience and meaning are modeled. It is precisely such contacts that are anticipated, and it is only such contacts that can falsify such anticipations. The general pragmatist conceptions of anticipation, then, cannot be defined except in terms of such "givens": without such contact, the constructions of anticipatory interpretations are completely free — there is nothing that can produce independent failure of anticipations — and the position falls into idealism. But the *epistemic character* of uninterpreted "givens," in turn, is encodingist, either by explicit model, or by neglect and default.

Representation as Function versus The Function of Representation. To note that the primary *function* of representation is anticipation of future potentialities is not the same as the interactive claim that the very *nature* of representation *is* that function of indication, of anticipation — the indication of future interactive potentialities. Representational content is fundamentally constituted as indications of interactive potentialities. Kelly's position, as did Peirce's, retains a classical conception of what representation *is* at the level of the "given" — however much it carries forward the pragmatic notions of what representation is *for* and the constructivist notions of how representation is *revised*.

Interactivism escapes this problematic because what is anticipated, what is indicated as potentiality, is not sense data or events or quale or any other sort of epistemic contact. To hold such a model is simply to re-encounter all the basic epistemic questions at that level of contact, or "given." Instead, the interactive model holds that what is indicated are potentialities of *interaction*, of flow of internal system process while engaged in interaction. Such a flow of *internal* system process is functionally and physically available in the system itself — available to falsify, or to not falsify, various goal switches that are functionally responsive to such internal conditions. This *functional* availability of internal conditions, most importantly, does not involve any *epistemic* contact with those internal conditions. The internal conditions and ensuing processes of the system are jointly *constitutive* of representation, but are not themselves representational. There is no epistemic "given" upon which the rest of the model is constructed, and with respect to which the anticipations and so on are defined. The epistemic relationship between the system and its environment is one of implicit definitions, and

internal indications among implicitly defining interactive procedures; the epistemic relationship is *not* one that involves any sort of epistemic inputs.

In many respects, pragmatism overlaps with and anticipates interactivism at least as much as any other position in the general literature. The remaining step, however, from anticipation as the *function* of representation to anticipation as the *ontology* of representation — the step that eliminates the epistemological "given" — is crucial because:

- It is only with this step that the logical confusions and aporias of encodingism are avoided.
- It is only with this step that representation is no longer construed as consisting of (or constructed upon) correspondences with the world resulting from the processing of inputs.
- It is only with this step that it is recognized that not only can passive systems not *revise* representations, passive systems cannot *have* any representations.
- It is only with this step that the necessity of timing is recognized not just for action and interaction, but for representation itself.
- It is only with this step that the grounding of representation in implicit definition and differentiation (not in correspondence) can be recognized.
- It is only with this step that the unboundedness of representation, as in the frame problems (see below), can be modeled.

The step from pragmatism to interactivism, then, is in one perspective 'just' a shift from "the function of representation is anticipation" to "representation is the function of anticipation." The consequences of that shift, however, are profound.

Failure to reach an internally indicated internal condition constitutes instrumental error — failure of pragmatic anticipation. Such pragmatic error will also constitute error of the predication that is implicit in that indication. That is *the* point of emergence of representational error out of pragmatic error, and, therefore, of representation out of pragmatics.

So, differences — important differences — remain between the general pragmatist notions of representation and the interactive model of representation. These differences, however, are minor compared to the gulf that exists between the interactive model and the dominant

encodingist — representation-as-correspondence — approaches in Artificial Intelligence and Cognitive Science. Ken Ford's pragmatist proposals concerning representation, then, are among the most compatible with the interactive approach to be found in contemporary literature.

Kuipers' Critters

Ben Kuipers and colleagues are exploring the design and development of *critters* — interactive robots that have no apriori interpretations of their inputs or their outputs (Kuipers, 1988; Kuipers & Byun, 1991; Pierce & Kuipers, 1990). Much of the work has focused on conditions in which the inputs are generated by various sensors equivalent to such senses as sonar and smell, and the outputs induce movement in a space, but, to reiterate, these interpretations are known *solely* from the designers' perspective, not from the critter's perspective. Two kinds of problems that have been investigated have been those of learning to navigate in spaces of hallways and walls, and learning to interact efficiently toward a goal.

Navigating Robots. One level of interest of this work concerns robots that can navigate, and that can learn to navigate, in real space. Thus, for example, inputs are deliberately errorful and so also are the consequences of outputs, and the critters must learn their way around in spite of this noise. Some of the strategies that have been developed include the recognition of places in terms of reproducibility of input patterns and sequences; hill climbing toward sensorily paradigmatic locations within neighborhoods, so that even errorful locomotion will be successful so long as it can get the critter into an appropriate neighborhood — at that point, the hill climbing can take over; the construction of topological maps of connectedness among places, and the ability to distinguish between different but sensorily indistinguishable places on the basis of their having different topological neighbors; and the building of metric information progressively into such topological maps.

No Prior Interpretations. From the interactive perspective, however, the most important aspect of this work is that the critters are fundamentally engaged in learning organizations of sensory-motor interactions with reactive environments. *Neither the inputs nor the outputs have any* ***intrinsic*** *meaning or interpretation for these systems*. The "fact" that the critters are running in (simulated) hallways affects only the sorts of input feedback that the critters receive in response to their outputs. The environment simulating programs create input to the

critters on the basis of a designer assumption of hallways and of outputs that run tractor treads, but the environments exist for the critters themselves *only* as organizations of potential interactions — organizations that they must explore and discover for themselves.

In any actual organism with a nervous system, there are oscillations traveling *in* along various axons and oscillations traveling *out* along various axons, and there are processes that internally relate them, *and that is epistemically* ***all*** *there is*. The entire familiar world must somehow be constructed out of *organizations* of potential interactions between such outputs and such inputs (Bickhard, 1980b; Bickhard & Richie, 1983; Piaget, 1954, 1971, 1977, 1985). The encoding presupposition that somehow the system already knows what is on the other end of its inputs, the assumption that the inputs *are* encodings, has utterly obscured and distorted this fundamental point.

Kuipers has understood this basic epistemological situation, and has shown how systems without the usual designer provided prescience can in fact, nevertheless, learn to successfully move around in a space — even though the basic cognitive capacities of the critters know nothing about movements or spaces per se. The critters functionally "know" only about how to reproduce various interactive flows, sequences, and patterns of outputs and inputs. In the terminology of interactivism, the critters construct a situation image, a functional representation of the organization of potential interactions, of their environments.

This is the epistemological position that all epistemic agents are in. Only the chimera of encodingism maintains the myth that inputs provide encodings of the system's environment. In Kuipers' critters, then, we find a foundational convergence with the interactive epistemology.

Extensions. The critters research programme is new, and, correspondingly, incomplete. Interactivism suggests some possible extensions which we would like to offer.

Situation Image. The first has to do with the potential representational power of the sensory-motor organization — situation image — form of representation. Learning to traverse a spatial environment is especially suited to this form of representation, and it might at first appear that interactive representation would be limited to more clearly sensory-motor knowledge such as this. When multiple sorts of interactions are taken into account, however — corresponding for example to such interactions as manipulations — and when richer

possible organizations in a situation image are considered — invariances of *patterns* of possible interactions for example — then the full representational capacity of the human child becomes in principle attainable (Bickhard, 1980b, 1992a, 1992c; Piaget, 1954). With a still richer architecture, abstract knowledge, such as of mathematics, becomes constructable (Campbell & Bickhard, 1986).

Goals as States. The second possible suggestion would shift the notion of goal from that of some input as a goal to be achieved (Pierce & Kuipers, 1990) to that of some system internal condition to be achieved. For the purposes of the research that was involved, this is a difference that makes little difference. But for more general problems, we suggest that such a shift offers considerable additional power. For example, multiple possible inputs, and multiple possible trajectories of input-output interactions, might all be equally successful in achieving some internal system condition goal, such as "blood sugar cells firing." An internal system-condition notion of goals abstracts from multiple paths of achievability; an input notion of goals cannot.

It might appear that what counts as an input could always in principle be redefined so as to take into account such possibilities as multiple paths of achievability — just define an input at whatever the final common path of system condition is — but this line of reasoning makes the incorrect assumption that characterizations of system organizations are superfluous relative to characterizations of their space of possible histories. Unless such a space of possible histories is strictly finite, finite and bounded numbers of histories of finite and bounded length, this is simply false — and very simple system organizations can involve infinite classes of possible histories (Hopcroft & Ullman, 1979; Minsky, 1967).

Goals defined in terms of internal system conditions also do not require any ad hoc differentiation of inputs with respect to those that are "just" inputs and those that are goals. Correspondingly, it does not require a fixed design of possible goals for a system, but makes possible the construction of new ones. As we argue elsewhere (see discussions of learning below), goals constitute a system's knowledge of what counts as error, and all a system can learn is at best how to avoid what it takes to be errors. Any general learning system, then, must have a flexible ability to construct new error criteria and to respond to them with learning trials.

Still further, goals as internal conditions do not pose epistemic problems — this is a *functionally* definable notion of goal. The system

does not have to recognize and discriminate represented goal conditions — such as a particular input — but, instead, "merely" to functionally respond to its own internal functional conditions in accordance with its own functional organization. It is this possibility that makes possible the construal of representation in terms of goals without involving a circularity. A goal in this sense is just a moderately complicated internal switch (Bickhard, 1993a). It is, of course, also possible for an internal goal condition to depend on a prior interactive determination that some external — represented — goal condition obtains. But this more complex epistemic version of goal is derivative from the strictly functional version: the functional notion of goal is required in order for the possibility of falsification of interactive predications — representations with truth values — to emerge out of non-epistemic, strictly functional, organization.

Dynamic Environments. A third extension has to do with the fact that the environments explored to this point in the critters programme have been static. The interactive perspective points out that extensions to more general environments will require, among other things, timing considerations inherent in the interactive knowledge. Such extensions will also require updating procedures for the situation image representations — they will require apperceptive procedures.

Learning. A fourth extension concerns the processes of learning per se. While the epistemology of critters has not involved any designer-provided prescience at all, the learning processes that they undertake *has* involved such designer prescience about the kind of environments that the critters have been expected to explore. As an initial step in a research programme, this is probably inevitable and necessary. But the implications of the interactive model are that general learning processes will have to be some form of variation and selection constructivism.

Still further, more remote, possible extensions are suggested by the general macro-evolutionary model underlying interactivism concerning the emergent nature and evolution of knowing, learning, emotions, and consciousness (Bickhard, 1980a; Campbell & Bickhard, 1986), and of the constructions of higher order processes such as self-scaffolding skills for learning, values, rationality, social participation, language, and the person (Bickhard, 1987, 1991d, 1992a, 1992b, 1992c, in press-a, in preparation-a, in preparation-b; Bickhard & Campbell, 1992). These, however, are obviously far in the future, highly programmatic, extensions.

The central point for our purposes is that Kuipers' critters constitute a realization and a demonstration of the basic interactivist epistemological position; they constitute a convergence at the level of that epistemological foundation. Consequently, they constitute an origin from which such extensions might begin to be explored. No encodingism model, no matter how complex, sophisticated, big, or useful, could possibly do the same.

Dynamic Systems Approaches

There has been a growing appreciation of dynamic systems approaches in recent years (Beer, in press-a, in press-b; Hooker, in preparation; Horgan & Tienson, 1992, 1993, 1994; Maes, 1990a, 1991, 1992, 1993, 1994; Malcolm, Smithers, & Hallam, 1989; Steels, 1994; Port & van Gelder, in press). The possibilities of analysis in terms of phase space dynamics, of open systems, of integrated system-environment models, and of non-linear dynamics — such as attractors, bifurcations, chaotic phenomena, emergent behavior, emergent self-organization, and so on — have excited a number of people with their potential modeling power. Dynamic systems approaches offer promise of being able to model — and to model naturalistically — many phenomena, including emergent phenomena, that have remained inexplicable within alternative approaches. Dynamic systems approaches can also model processes that are uncomputable in standard frameworks (Horgan & Tienson, 1992, 1993, 1994). The interactive model shares the goal of a thorough and consistent naturalism, and arises in the framework of dynamic open systems analysis (Bickhard, 1973, 1980a, 1980b, 1992c, 1993a; Bickhard & D. Campbell, in preparation). There are, therefore, important convergences with much of the work in this area. There are also, as might be expected, differences — especially with respect to issues of representation per se.

Cliff Hooker. Hooker (in preparation) proposes a naturalist philosophy of the natural sciences of intelligent systems. He advocates a process theoretical approach — specifically, the organizations of processes that constitute complex adaptive self-organizing systems — toward a naturalized theory of intelligence. Hooker, Penfold, & Evans (1992) present a novel architectural approach to control theory — local vector (LV) control — and explore issues of problem solving, concepts, and conceptual structure from within that framework. Their aim is to

show how LV control "may be able to illuminate some aspects of cognitive science" (p. 71).

Hooker (1994) develops a model of rational thought within a dynamic systems framework. He advocates a consistent naturalism — including a naturalism of rationality — and explores the contributions toward, and errors with respect to, that goal in literature ranging through Popper, Rescher, and Piaget. In the course of this analysis, Hooker develops a model of a non-foundationalist evolutionary epistemology (D. Campbell, 1974). His discussion convincingly demonstrates the plausibility of a naturalism of rationality. Hooker's is a truly kindred programme.

Concerning naturalism, for example: If mind is *not* a natural part of the natural world, then Artificial Intelligence and Cognitive Science have set themselves an impossible task. The commonalities regarding dynamic, interactive, systems as the proper framework to explore are clear. Although rationality per se is not much discussed in this book, the interactive model of rationality emphasizes knowledge of possible errors (Bickhard, 1991d, in preparation-a), while Hooker emphasizes successful autoregulations. Ultimately, both error knowledge and autoregulation knowledge must function together; in many ways, these are complementary projects.

Tim van Gelder. Van Gelder (in press-a) points out that, from a dynamic systems perspective, standard computationalism consists of a severely limited subspace of possible kinds of dynamics — and a space of limited power relative to the whole space. He advocates moving from these limited visions of possible systems to a consideration of how input-output relationships, and feedback relationships, might be best modeled or best accomplished within the entire space of dynamic possibilities. The emphasis on general system dynamics and on feedback is strongly compatible with the interactive framework, but van Gelder, like Brooks (1991a) and others, avoids issues of representation. Any complex interacting system, however, will have to contend with multiple possible next interactions, and will have to select among such possibilities on the basis of indicated outcomes. Appropriate forms of such indications of interactive possibilities *constitute* interactive representation. We argue that ignoring such emergent possibilities within system organization can only weaken the overall approach: it ignores the power of representation.

Randall Beer. Beer (1990, in press-a, in press-b; Beer, Chiel, & Sterling, 1990) proposes adaptive systems as a framework for analysis

and design of intelligent systems. He proposes, in fact, the notion that intelligence *is* adaptive behavior. If so, then the design and understanding of intelligent systems must recognize them as being intrinsically embodied as some sort of agent that, in turn, is intrinsically embedded in an environment. Embodiedness is required for a system to be even potentially adaptive in its behavior, and environmental embeddedness is required for the notion of adaptiveness to make any sense.

Autonomous symbol manipulation is not an appropriate framework for such analyses. The symbol manipulation framework for Artificial Intelligence and Cognitive Science is neither embodied nor embedded. In fact, as discussed above, standard frameworks cannot account for any epistemic contact with an external world at all (Bickhard, 1993a; Fodor, 1981a; Shanon, 1993). Beer, accordingly, eschews standard representations.

Instead, he begins with relatively simple versions of adaptive systems — insects, in this case — and attempts to capture some of their interesting and adaptive behavior in artificial creatures. Specifically, investigations have addressed such phenomena as locomotion, exploration, feeding, and selections among behaviors. The focus is on the physical structure of the simulated insect agent, and on the artificial nervous system that underlies its dynamics of interaction. Beer makes "direct use of behavioral and neurological ideas from simpler animals to construct artificial nervous systems for controlling the behavior of autonomous agents." (1990, p. xvi). He has dubbed this approach "computational neuroethology" (see also Cliff, 1991). The rationale for computational neuroethology is a strong one: Nature has usually been smarter than theorists.

In further development of the computational neuroethology approach, Beer & Gallagher (1992; Gallagher & Beer, 1993) have demonstrated the evolution in neural nets, via a genetic algorithm, of the ability to engage in chemotaxis — movement toward a chemical concentration — and to control a six legged insect walker. Note, these are not just nets that can control chemotaxis and six-legged walking per se, but nets that *develop* that ability. Yamauchi & Beer (1994) present a similar ability for a net to learn to control sequences of outputs, and to switch between appropriate sequences when conditions of adaptiveness change.

On a more programmatic level, Beer (in press-a) argues against computationalism, including against correspondence notions of

representation. He urges dynamic systems as an alternative programme, but with no focus on issues of representation in dynamic systems. Beer (in press-b) addresses a wide range of programmatic issues. He discusses the importance of embodiment for understanding and designing agents, in contrast to the disembodied agent perspective that is typical of Artificial Intelligence, and of the inherent dynamic coupling of an embodied system with its environment (Bickhard, 1973, 1980a). He provides an introduction to the notions of dynamic systems analysis; provides an overview of the insect walker; and argues again against the notion that internal states, even internal states correlated with something in the environment, constitute representations.

Beer (in press-b) also raises an issue that we consider to be fundamental, but that is seldom addressed. What constitutes system survival in dynamic systems terms? The question is intimately connected to issues of how to model adaptiveness, and, therefore, of how to model the emergence of function and representation — notions that derive from survival and adaptiveness respectively (Bickhard, 1993a; Bickhard & D. Campbell, in preparation). Beer's answer is, to a first approximation, that persistence — survival — of a system is equivalent to the persistence of crucial limit cycles in its dynamics. He then proposes that this conception might be generalized in terms of the notion of autopoiesis (Maturana & Varela, 1980, 1987). These conceptualizations of survival and adaptiveness are strikingly similar to those in Bickhard (1973, 1980a), where a notion of system stability is defined in terms of the persistence of a condition of the reachability of system states. It is not clear that the conceptualizations are fully equivalent, but the first focus should be on the importance of the questions rather than on the details of the proposed answers.

Try Simpler Problems First. Complex problems, such as that of intelligence, are often broken down into simpler problems, or are addressed in terms of simpler versions of the problem, in order to facilitate investigation. The notion of "simpler problem," however, is quite different in a dynamic systems approach than in standard frameworks for investigating intelligence. Micro-worlds, restricted knowledge domains, chess playing, and so on, are examples of what counts as simpler problems from a symbol manipulation perspective. They are "simpler" in the sense that they do not commit to the presumed full symbol store and manipulation rules of a presumed fully intelligent entity. They are partial symbol manipulation "intelligences" in the sense

that they involve partial symbol domains and partial sets of manipulation rules.

But none of these are agents at all, in any sense of the notion of agent. And most certainly they are not adaptive agents. These approaches suppose that they are cutting nature at its joints — joints of symbol encoding domains — but, if intelligence cannot be understood except in adaptive agent terms, then such attempts have utterly missed the natural joints of intelligence. They are putting a giant buzz-saw through the house of intelligence, severing a piece of the living room together with a piece of the kitchen, and taking bits of chairs, tables, and other furniture along with it (or worse, a buzz-saw through the wood pile out back). That is no way to understand houses — or intelligence.

More carefully, encodingism is studying the placement, configuration, and color patterns of — and how to operate — the switches and controls on all the appliances and fixtures in the house. It misses entirely the manner in which stoves, faucets, air-conditioners, television sets, incandescent bulbs, and so on function internally, and at best hints at the more general functions such as cooking and eating, plumbing, lighting, and climate control that are being served. Encodingism provides not a clue about the existence or nature of electricity or water flow or air currents and properties, and so on. Encodingism is a static model of representation. Encodings can be manipulated in time, but an encoding relationship per se is atemporal — just like a light switch can be manipulated in time, but a switching relationship per se is atemporal. The crucial aspects are the functional dynamic aspects, and encodingism does not even address them. This is still no way to understand houses, or representation — or intelligence.

If intelligence is adaptive behavior, then "simpler" means simpler versions of adaptive agents and their behavior (Brooks, 1991c; Maes, 1993). The natural joints do not cut *across* the interactive dynamics between agent and environment, but, instead, divide out simpler *versions* of such dynamics — simpler versions such as Beer's artificial insects, or Brooks' robots, or Kuipers' critters.

Representation? Rather clearly, we are in fundamental agreement with Beer's dynamic and adaptive systems approach (Bickhard, 1973, 1980a, 1993a; Bickhard & D. Campbell, in preparation). For many reasons, this is a necessary framework for understanding mind and intelligence. We are also in agreement with his rejection of symbol manipulation architectures as forming a satisfactory encompassing

framework for that understanding. Beer also notes that the internal dynamics of his insect do not match the connectionist notions of representations. Although we regard connectionism as a potential, though currently somewhat misguided, ally, the misguidedness is precisely in the notion of representation that connectionists have adopted (see discussion below), so, again, we are in full agreement with Beer.

In focusing on and correctly rejecting contemporary notions of representation, however, Beer has overlooked the possibility that there might be a more acceptable, and compatible, model of representation. We argue that the interactive model of representation emerges naturally from dynamic adaptive systems perspectives (see below, and Bickhard, 1993a; Bickhard & D. Campbell, in preparation), and is not subject to the myriads of fatal problems of encodingism — whether in symbol manipulation or in connectionist versions. If so, then representation — properly understood — has a central role to play in analyzing and designing intelligent systems.

Robotic Agents. The move of roboticists toward systems dynamics more general than the classic input-processor-output architectures is becoming more and more widespread (Brooks, 1991b; Maes, 1990a, 1993; Malcolm, Smithers, & Hallam, 1989). We have discussed Brooks subsumption architectures, and his associated rejection of the notion of representation, and Beer's insects and neuroethology. The possibility — and the possible usefulness — of representation, however, remain a point of difference among roboticists.

Adaptive Agents. In general, approaches to the design of adaptive autonomous agents de-emphasize representation (Maes, 1993, 1994). There is "little emphasis on modeling the environment" and "no central representation shared by the several [functional] modules" (Maes, 1994, p. 142). In fact, there tends to be a de-centralization of control into multiple interacting functional modules, with at best very local "representations" resident in each module sufficient to the tasks of each module. Brooks' theme of "the environment is its own best model" is emphasized, and what "representations" do occur are much less likely to have the "propositional, objective, and declarative nature" typical of classical Artificial Intelligence. Maes (1993, 1994) provide overviews of the field; Maes (1990a) provides a number of examples.

One major approach emphasizes the close relationships between adaptive agents in the broad sense and living systems. This emphasis is often in terms of the behavioral problems that such agents encounter,

whether living or designed, and in terms of inspiration from solutions found in living systems for subsequent artificial design (Beer, 1990; Cliff, 1991; Steels, 1994). In this view, general models of adaptive autonomous agents constitute a framework for approaching the design of artificial intelligence via the design of artificial life (Steels, 1994). As is by now familiar, issues of representation are usually either rejected or ignored, though almost always in terms of correspondence notions of representation. But questions concerning the adequacy of dynamic systems frameworks to issues of cognition and representation cannot be postponed forever (Steels, 1994).

Exceptions do exist in which standard notions of representation are rejected, but, instead of concluding thereby that representation plays no role, alternative conceptions of representation are adumbrated. In keeping with the focus on interactive robots and open living systems, these alternative notions of representation tend to have a strong interactive flavor. Nevertheless, specifics of the distinction between epistemic contact and representational content, implicit definition and predication, emergent system detectable truth value, and so on, are absent. Intuitions that knowledge is fundamentally a matter of ability to maintain system integrity can be strong (Patel & Schnepf, 1992; Stewart, 1992 — see Bickhard, 1973, 1980a, 1993; Bickhard & D. Campbell, in preparation), but where to go from there is not so clear. Proposals can end up with some more complex version of input processing correspondences, in spite of the basic intuitions.

Tim Smithers. Malcolm & Smithers (1990) actively explore novel and hybrid architectures, with a dynamic systems framework as an underlying inspiration. Notions of representation are based on interactive robotic submodules, but interpretation of representations is still primarily from the perspective of an observer — a designer or user. Smithers (1992) rejects folk psychology models of intelligence as being not useful in the design of intelligent systems — notions of belief and desire have demonstrated themselves to be unhelpful. He advocates a dynamical systems approach instead. Smithers (1994) refines such a dynamical systems approach, proposing that agent-environment systems be approached within the conceptual framework of nonlinear dissipative dynamical systems theory. Smithers points out the necessity for considering situated, embodied agents — and agents for which history counts — and outlines a systems approach for doing just that.

Nehmzow & Smithers (1991, 1992) present robots that self-organize maps based on their sensory-motor experience. Under certain conditions, these maps will begin to capture the basic physical layout of the environment in terms of its interactive potentialities. But to do so requires progressively finer differentiations with respect to the robot's interactions so that those differentiations succeed in discriminating differing physical locations. With inadequately fine differentiations, for example, two corners that "look alike" in immediate sensory-motor interactions will not be discriminated. This process of differentiating the environment in terms of interaction paths and histories, and the future "histories" that those past histories indicate as being possible, is the fundamental form of interactive representation. (See also Kuipers, 1988; Kuipers & Byun, 1991; Mataric, 1991.)

Emergent Action Selection. One of the important themes of this orientation is the emergence of behavior from multiple loci of control (Steels, 1991, 1994; Maes, 1993, 1994). There need not be any central controller or planner for adaptive behavior to occur (Beer, 1990; Brooks, 1991a; Cherian & Troxell, 1994a, 1994b, in press; Cliff, 1992; Pfeifer & Verschure, 1992). Multiple interacting and competing sources of behavioral control can generate complex adaptive behavior emergent from those interactions and competitions.

Maes (1991, 1992) describes an architecture that can emergently select actions and action sequences appropriate to various goals, and that can learn such selections. This moves into the critical realms of motivation and learning. As discussed in the section on Benny Shanon, motivation as action selection is an intrinsic aspect of the interactive model — issues of motivational selections and issues of representation cannot be separated except as aspects of one underlying process organization.

Wade Troxell and Sunil Cherian. Again, in such multiple-module action selection, representation, at least in its classical symbol manipulation sense, is de-emphasized or rejected. "No explicit, isomorphic mapping is performed from objects and events in some real external world to internal system states. The view that 'intelligence is the process of representation manipulation' is rejected in favor of strategies that place more emphasis on the dynamics of task-directed agent-environment interactions." (Cherian & Troxell, 1994a).

If representation is an emergent of interactive system organization, however, as the interactive model argues, then robotics, and dynamic

systems investigations more broadly, are the natural fields in which issues of representation ultimately cannot be avoided, and in which genuine representations will be naturally emergent. Representational design issues are robotics issues, not computational issues more narrowly (Bickhard, 1982). Questions of representation can be *raised*, and have been addressed voluminously, within a computational or a connectionist framework, but they cannot be *answered* without moving to a full interactive system framework.

Cherian & Troxell (in press) propose just such an interactive framework for exploring issues of knowledge and representation in autonomous systems. They reject a structural encoding model of knowledge in favor of a notion of knowledge as the ability to engage in successful interaction with an environment. In this interactive sense, knowledge is constituted in the organization of the system's control structure, not in static encoding structures. Knowledge as constituted in interactive control structures does not require an interpreter of encodings; therefore, it does not encounter the fatal aporia of encodingism. They offer a formal description for interactive control structures, and an example application.

Modeling knowledge and representation as properties of interactive competence is inherently natural for robotics, once the mesmerizing seductions of encodingism are overcome. We argue that it is the only viable approach not only for robotics per se, but for Artificial Intelligence and Cognitive Science more broadly — and for psychology and philosophy. Embodied autonomous interactive dynamic systems are not just an application aspiration of Artificial Intelligence, they are the essential locus of emergence, understanding, and design, of representation, knowledge, and intelligence — of mind. They are the locus of resolution of the fundamental *programmatic* aspirations of Artificial Intelligence and Cognitive Science.

Representation versus Non-representation. The debate about the role of representation in dynamic systems approaches has construed representation in the classical encodingist sense. The issues debated are whether dynamic systems do, or need to, involve internal states that track external objects, events, properties, and so on, such that that tracking is an important aspect of the explanation of the functioning of the system (Cherian & Troxell, in press, and Hooker, in press, are rare exceptions). Some argue that the dynamic couplings between system and environment are such, or can be such, that representational construals are not useful.

Others argue that they *are* useful (e.g., Horgan & Tienson, 1992, 1993, 1994). Clark and Toribio (in press) point out that representational construals are to be expected to be important primarily in "representation hungry" problem domains, by which is meant:

> 1) The problem involves reasoning about absent, non-existent, or counterfactual states of affairs.
> 2) The problem requires the agent to be selectively sensitive to parameters whose ambient physical manifestations are complex and unruly (for example, open-endedly disjunctive). (p. 16, manuscript)

These conditions virtually require that something in the system track the relevant states of affairs or parameters. We agree that such "representation hungry" problem domains exist, and that they will generally require environmental tracking of the sort that Clark and Toribio argue for.

Such tracking, however, can be accomplished in ways that do not look much like standard notions of elements-in-representational-correspondences. The tracking of correspondences can be accomplished not only by states and limit cycles, but also by hyperplanes or even more general manifolds in the dynamic phase space. The system entering one of several alternative such manifolds could constitute the tracking or "recording" of some condition — a correspondence between dynamic manifold and condition rather than state and condition. It is not clear that such "tracking by manifold" satisfies standard intuitions about representation.[13]

More deeply, however, we disagree that *any* such tracking will constitute representation. Interestingly, this disagreement extends to both sides of the issue, because those arguing *against* the usefulness of representation in dynamic systems approaches use much the same notion of representation.

At times, additional restrictions are assumed by those arguing against representation, such as that the representations be some version of standard explicit syntactic symbols. With such a notion of representation, it is correct that dynamic systems approaches need not necessarily involve

---

[13] This is the continuous phase space equivalent of state-splitting in finite state machines (Hartmanis & Stearns, 1966; Booth, 1968; Ginzburg, 1968; Brainerd & Landweber, 1974; Bickhard, 1980b; Bickhard & Richie, 1983). In either case, the functional consequences of any purported "representations" are absorbed into the dynamics of the system, leaving little left to "be" the alleged representations.

symbolic representation, at least in most cases. Caveats *might* be necessary for the arguments, for example, concerning systematicity in thought (Fodor & Pylyshyn, 1981; see, however, Clark, 1993; Hooker, in preparation; Niklasson & van Gelder, 1994; Port & van Gelder, in press; and the discussions of connectionism below). This stance, exemplified by many of the researchers in dynamic systems approaches, that symbol manipulation approaches are not needed or are even detrimental, constitutes a clear rejection of the classical symbol manipulation framework — a rejection with which we are strongly in agreement. The tracking notions of representation, however, are weaker models of representation, and, therefore, stronger models upon which to base a claim of representation in dynamic systems (cf. Touretzky & Pomerleau, 1994; Vera & Simon, 1994).

Representations as Correspondences? The reasons for our disagreement with this assumption (common to both sides of the representational versus non-representational argument), that representations are encoding correspondences, is by now obvious. We argue that such tracking states do *not* constitute representations, so, even if they *are* required in some conditions, that still does not amount to a requirement for representation. Clark and Toribio (in press) and Clark (1994) argue for such tracking notions of representation as constituting the best model of representation available. They acknowledge that this provides at best an "external" notion of representation, dependent on the interpretations and explanations of observers and analyzers, but claim that "internal" notions of representation fall to homunculus problems — they require internal homunculi to interpret the internal representations. We agree with this criticism in general, but not with the assumption that there is no alternative notion of internal representation that is not subject to this criticism.

Observer Representations. "External" notions of representation become rather unsatisfactory — the intrinsic circularity or infinite regress become obvious — when considering the representations of those external observers themselves. "External" encodingism avoids *internal* interpretive homunculi only at the cost of *external* interpretive homunculi, and the promissory note issued by those homunculi cannot be cashed externally any more than internally. The fact that there *do* exist some external interpretive homunculi — people, for example — unlike the *non*-existence of internal interpretive homunculi, does nothing toward *accounting* for those external interpreters. Accounting for human

intentionality and intelligence was the problem to be addressed in the first place, and it cannot be addressed simply by reference to other — "external" — intentional and intelligent agents.

Conditions that Require Representations. On the other hand, we would argue that there *are* problem conditions and system conditions that do require representations — even in dynamical systems, even in the full interactive sense of representation. These conditions follow directly from the interactive model itself. If a system faces more than one possible next action or course of interaction, and must select among them on the basis of indicated internal outcomes of those interactions, then those indications constitute interactive representation — capable of being false and falsifiable internal to the system itself.

Not all interaction selections will necessarily be based on indications of the internal outcomes of the to-be-selected interactions. If the environment is sufficiently anticipatible, and the system requirements sufficiently stable, then interaction selections might be simple evocations — e.g., triggerings — by internal system state: In certain internal system conditions *this* interaction is selected, while in other conditions *that* interaction is selected, with no need for indications of subsequent internal outcomes. If the selections of interactions can be strictly feed-forward and informationally ballistic (Bickhard, 1980b) in that sense, then interactive representation need not be present.

This is so even if those selections are based on internal states that factually track external conditions. Such tracking is a functional and factual state of affairs, one that may even be necessary to the survival of the system, but note — once again — that nothing about the existence or success or failure of that tracking per se is available to the system itself. If there is no possibility of system representational error, then there is no system representation.

Interactive representation is required, then, when the processing in the system must be potentially controllable, at least in part, by system error in achieving its indicated internal outcomes. This will occur when the conditions and interactions that will yield success or failure are uncertain, and therefore the possibility of such error must be functionally taken into account — for example, in goal-directed systems.

Such uncertainty of outcome, in turn, could result from a randomness of environmental response, or a complexity of relevant environmental conditions that is too great to be detectable in reasonable time (and other resource costs). Another source of such uncertainty

would be a system that is sufficiently complex that many of its interactions are novel for the system. For such a system, even if the environment were stable and simple enough that ballistic actions with no error feedback would be possible in principle, the system would nevertheless not be able to anticipate its interaction outcomes because it would not have had sufficient prior experience with its novel interactions, and would, therefore, have to be able to take error into account. In other words, the uncertainty that would require outcome indications — interactive representation — can be a property of the environment, or a property of the system's interactive knowledge about that environment.

System generated error is required when system implicit anticipations of the courses and outcomes of interactions cannot be assured. Of course, system generated error, once available, might be useful and used for many conditions in which it is not strictly necessary, but is less costly than alternatives, as well as in conditions in which it is required.

One critically important version of system error guided processes, of course, is that of goal-directed interactions (Bickhard, 1980b). Another is that of general learning processes. Learning cannot be fully successfully anticipatory — if it were, there would be nothing to be learned. Learning must involve the possibility of error, and such error must be functionally detectable by the system itself so that the learning can be guided by it (Bickhard, 1980a, 1992a, 1993a, in preparation-c; Campbell & Bickhard, 1986; Drescher, 1991 — see the discussion of learning below).

So, in problem domains that involve sufficient uncertainty of environmental response to — and, therefore, of system outcome of — system-environment interactions, system detectable error can be necessary. System detectable error of anticipated (indicated) system internal outcomes *is* interactive representation. Dynamic system approaches to system-environment couplings that involve uncertainty of the course of the interactions, then, can require interactive representation in the system in order to be competent to functionally respond to the inevitable error.

The Emergence of Function. This discussion, of course, relies on a notion of something counting as error *for a system*. Internal indications of the internal outcomes of interactions provide a system detectable condition of whether or not those indicated conditions obtain, but what is to count as an emergence of success and failure here? Why would

achievement of indicated conditions, for example, count as success rather than failure? Or, under what conditions would such achievement count as success and under what conditions as failure?

The notion of failure here is that of functional failure. The representational emergence involved is out of a functional level of analysis. But functional failure too must be naturalized — it too must be naturally emergent, with no dependence on the external interpretations of observers, users, and so on. There is a rich literature of approaches and problems concerning the naturalization of such notions of function and dysfunction (e.g., Bechtel, 1986; Bigelow & Pargetter, 1987; Boorse, 1976; Cummins, 1975; Dretske, 1988; Millikan, 1984, 1993; Neander, 1991; Wimsatt, 1972; Wright, 1973).

We propose to derive a model of emergent function in a framework of open dynamic systems. Open systems require interaction with their environments in order to continue to exist; they require continuing interchange of matter and energy. Complex open systems, especially complexities involving selections among alternatives for the activity of the system, can contribute toward or harm the continued existence of an open system via contributions toward or harm to the environmental conditions or system-environmental relationships upon which the existence of the system is dependent.

A too simple example is a flame, which contributes to the maintenance of the threshold temperature necessary for the flame's continued existence — and thereby also to the maintenance of the supply of oxygen via convection of air. This example is too simple in that there are no selections on the part of the flame among alternative manners of contributing toward its self-maintenance.

Even the simplest living systems, however, manifest internal homeostasis maintained by selections among alternative processes in those systems. A bacterium's selections of tumbling or swimming, in "getting worse" and "getting better" environments respectively, is one such example (D. Campbell, 1974, 1990). Here, swimming will halt and tumbling will ensue if, for example, the bacterium is swimming *down* a sugar gradient, thus making things worse, but swimming will continue if swimming *up* a sugar gradient, thus making things better. The selections of tumbling or swimming constitute environmentally sensitive selections among alternative actions — selections for those actions that contribute to the survival of the system in those respective environments.

The *contributions* of such selections toward system survival constitute *functions* of those selections in a sense of function that does not require any external observer to notice them. The natural reality of such functions is manifested in the continued existence, or failure of existence, of the system itself — and of the natural, causal, consequences of that existence or dissolution. Such contributions of selections toward system existence grounds a naturalistic emergent notion of function, upon which a full framework of functional analysis can be developed — in particular, a framework supporting the modeling of emergent interactive representation. A more detailed elaboration of this analysis of the emergence of function is presented elsewhere (Bickhard, 1993a).

The interactive model, then, connects deeply with dynamic systems approaches: open interactive dynamic systems are *required* for the naturalistic emergence of the critical phenomena of function and representation. The interactive model of the nature of such representation, in turn, requires representation in conditions of uncertainty in system-environment interactions, thus taking a stance in the representation versus non-representation debate concerning dynamic system approaches. The interactive position that representation is required in dynamic systems approaches, however, involves a rejection of the notions of representation that are held in common to both sides of this argument. Strictly, then, we end up with a rejection of the terms of the dispute and a claim to transcend that dispute.

The emphasis, however, should remain on the sense in which dynamic systems approaches are not only very powerful modeling tools, but are *necessary* for understanding function and representation in naturalistic terms. It is also worth re-emphasizing that the interactive dynamic systems approach is not only necessary for understanding *representation*, but that it has strong claims even in the heartland of the symbol manipulation approach — language (Bickhard, 1980b, 1987, 1992a, 1992c, in press-a; Bickhard & Campbell, 1992; Campbell & Bickhard, 1992a) Interactive dynamic systems approaches are the frontier for future exploration of cognition.

A DIAGNOSIS OF THE FRAME PROBLEMS: IMPLICIT REPRESENTATION, EXPLICIT REPRESENTATION, AND PROLIFERATION

Any attempt to capture the representational power of an interactive system within an encodingist framework encounters fatal problems of representational proliferation (Bickhard, 1980b). There are several ways in which the impossibility of such a replacement manifests itself, and we wish to explore some of these and to argue that they include the general class of problems in Artificial Intelligence known as the frame problems (Pylyshyn, 1987; Ford & Hayes, 1991; Genesereth & Nilsson, 1987; Toth, in press).

The original frame problem (McCarthy & Hayes, 1969) was focused on how an Artificial Intelligence system could represent and make use of common sense knowledge about what does and does not change as a result of actions in the world. The problem emerges when it is recognized that the relevancies, or lack thereof, of some action to the multiple parts and aspects of the world is not determinable a priori (Glymour, 1987). Instead, any particular assumptions about such relevancies, or their absence, seem to be defeatable by appropriate constructions of contexts or histories. If I move this book to the other room, for example, its color will not change — unless there is a paint shower in the path through which I move it. If I move this book, the house will not blow up — unless there is a bomb connected to a pressure switch under the book — unless the bomb is defective. And so on. Attempting to represent all such relevancies, lack of relevancies, and their iterated defeating conditions *seems* impossible. Yet people do something like that all the time.

A class of problems of computation and representation that seem related to the frame problem have grown up, and some of them are sometimes called “the” or “a” frame problem. There is, in fact, no consensus about exactly what the frame problem is, and some degree of contention concerning who has the legitimate authority to adjudicate among the contenders (Pylyshyn, 1987; Ford & Hayes, 1991). We are not so concerned about the pedigree of various versions of and relations among the frame problems, but rather with a class of representational and computational problems that emerge in attempting to render an interactive system in an encodingist architecture. These problems seem to include most of the various frame problems; if so, this analysis should at least

provide a kind of diagnosis of the frame problems, and perhaps some suggestions about its "solution" — or dissolution.

We begin with some general observations about interactive representation, how it differs from encoding representation, and how those differences manifest themselves as proliferations, even unbounded proliferations, of encodings. Many of these initial differences are themselves aspects of still deeper differences between interactivism and encodingism, and we end with explorations of some of those more fundamental differences and their consequences.

Some Interactivism-Encodingism Differences

Our first observation is that interactive representation of the current situation is constituted as indications concerning potential interactions in the situation image. That is, the situation image is constituted out of indications of interactive *relevancies* — relevancies of completing some interactions for the possibilities of completing other interactions. Representation in the situation image is intrinsically relational — indicative relationships between potential interaction outcomes and further potential interactions (Bickhard, 1980b).

To render such intrinsically relational representation in terms of intrinsically **in**dependent encoding atoms destroys the web of relational relevancies, and requires that they somehow be built back in to the representation as distinct encodings. That is, to move from intrinsically *relational* functional indications to independent encoding *elements* destroys the relevancy information, the relational information.

It might seem that that relational information could itself be encoded — for example, as relational terms in a language, with expressions using those terms to relate various elements. But there is no a priori solution to how to do this, and all the relational relevancy information has been destroyed in atomizing it; all encodings might conceivably be relevant to all others, and each case (or type of case) requires a separate encoding of its actual relevance or lack thereof. Any changes, such as the result of an action, requires that all *possible* such relevancies be either taken into account with concomitant changes in ramified consequences, or taken as not actual relevancies. This is at least combinatorially explosive, with consequent memory and computational impossibilities, if not unbounded.

Note secondly that these relevancies in the situation image are between outcomes of interaction *types*, specified by the procedures that

engage in those interactions. These constitute differentiators, and they implicitly define categories of situations that yield those outcomes should they be interacted with. Such implicit definitions are open, deictic, indexical, and context dependent. Among other consequences, there is no explicit representation of what is implicitly defined, and, therefore, no knowledge of how many instances, versions, variations, or subtypes there are of what is implicitly defined.

The apparently simple problem of examining the instances of a given differentiation type — the extension set of a differentiator — is, therefore, uncomputable. It is uncomputable even if the extension set is assumed to be finite, so long as it is not assumed that the cardinality of the extension set is known or has a known upper bound. There can be no assurance that that set has been exhausted in any finite search — there might always be one more instance not yet examined. Attempting to capture this within the typical AI Cartesian world, in which each object has a unique name and location, requires either moving to a toy world in which the differentiated sets are known to be finite, and explicitly enumerated, via the God-like surview of the designer of the system, or it requires an infinite number of names together with a *truly* God-like surview of the universe. The typical approach requires an assumption of *omniscience* (Toth, in press).

As Agre (1988) points out, presumption of such a naming encoding also requires special and inefficient machinery for generalizations over named instances in learning processes; otherwise, even if the program learns how to put block A on block B, it would have to learn independently how to put block C on block D, and so on. The implicitness of differentiators offers such generalizations intrinsically: putting block-in-front on top of block-to-the-side works no matter what the names of the blocks, including no names at all. Implicitly defined sets can be generalized over, are *intrinsically* generalized over, without enumeration.

A third observation concerns the relationship between the situation image indications that are actually constructed at any given time, and those that *could* be constructed on the basis of those already present if apperceptive computation were not itself subject to limitation. The organization of indications that have been explicitly constructed is called the explicit situation image, while the organization that could be constructed with unlimited computational resources is called the implicit situation image (Bickhard, 1980b). The basic point to be made at this

time is that the apperceptive procedures, the procedures that do construct the explicit situation image and that would construct the rest of the implicit situation image if there were no resource limits, *implicitly define* the class of potential situation image indications in the implicit situation image. The implicit situation image is, in spite of its being largely implicit, the strongest candidate for the system's representation of its situation. Further explicit indications can be constructed as needed, unless resource costs interfere. Just as the implicit definitions of differentiators capture an important and real aspect of system knowledge, in spite of the open implicitness, so also do the implicit potentialities of apperceptive procedures.

For still another instance, note that the interactions that are indicated as possible in the situation image are, strictly, classes of possible interactions, *types* of possible interactions, distinguished by the procedures that would engage in them and, perhaps, by the particular final states that are designated for them. That is, for interactive representation, what is implicitly predicated of an implicitly defined and differentiated situation is the possibility of an implicitly defined class of actual interactions. The ubiquitousness of implicitness in interactive representation should now be obvious; we will also argue that it is of fundamental importance (Dennett 1987; Dreyfus, 1981).

Implicit versus Explicit Classes of Input Strings

Consider the combinatoric space generated by a set of atomic encodings. In general, subsets of that space, which consist of sets of encoding strings, are not equivalent to any single string in the space. Exceptions can exist for finite such sets, which might be considered (for some purposes) to be equivalent to the string obtained by concatenating the strings in the set in some particular order. Even in these cases, it is a rather strained equivalence, and most of these subsets, in any case, will be unbounded. Furthermore, no correspondence can be defined between those subsets and the strings in the space: a power set is larger than its base set.

A general subset of encoding strings, then, can***not*** be represented by any single such encoding string. Insofar as the encoding strings are taken to be encoding representations themselves, these subsets of possible such strings lie outside of the representational power of that encoding space. This point is general to any such space generated by any set of atomic encodings.

A new *atomic* encoding might be *defined* for any particular such subset of possible encodings, should this be desired, but, as a general strategy, this would require a very large infinity of new atomic encodings as the power sets were iterated up the cardinals. Furthermore, there is no way for any such new ad hoc encoding to be given the requisite representational power. It cannot stand-in for any prior encoding string, since there is none that will represent the subset, and the intuition — or stipulation — that it should represent that subset as a whole is dependent on the *observer* of the set theory. So, even if it is presupposed, contrary to fact, that there could be some observer independent base of atomic encodings, the encoding of *subsets of strings* in the generated encoding space will be necessarily derivative from an observer, and impossible in general in any case because of the infinities involved.

Recognizers. Strings in such subsets of strings, however, can be recognized by an automata theoretic recognizer — an automaton with specified start state and final states is said to recognize an input string if that string leaves the automaton in one of its final states at the termination of the string (Brainerd & Landweber, 1974; Hopcroft & Ullman, 1979). Such a recognizer will *implicitly define* the *class* of such strings that it can recognize. This, in fact, was the paradigmatic form in which the interactive notion of implicit definition was first explicated. This move to automata abstracts away from any genuine outputs and interactions, and from all timing considerations, of the general interactive model. The basic idea of such implicit definition can be generalized to more powerful abstract machines (Hopcroft & Ullman, 1979).

This interactive notion of implicit definition is related to, and derivative from, the sense in which a class of formal sentences will implicitly define the class of possible models for those sentences — but it is not the same. The interactiveness of this notion, for one difference, renders it far more powerful as a means of differentiation than the atemporal, arbitrary, point to point mappings of model theory (Demopoulos & Friedman, 1989).

It should also be pointed out that even the formal model-theoretic sense of implicit definition is *at least* as powerful as explicit definition (Quine, 1966a) — implicit definition is not an enervated weak version even in its formal sense. Implicit definition is in fact *more* powerful when the consideration is taken into account that explicit definition requires other symbols out of which such explicit definitions can be constructed, and that such other symbols are not in general available. In

particular, they are not available for the construction of *any* novel representations, representations that are not simply combinations of representations already available. Explicit definition, in fact, is clearly at best the first step of the regress of symbols defined in terms of symbols which are defined in terms of symbols, and so on — a regress that cannot be halted merely with still more symbols. Implicit definition does not require any such further symbols, or any other form of further representation. Implicit definition does not require representation in order to construct representation (Bickhard, 1993a; Bickhard & Campbell 1992; Campbell & Bickhard, 1992a).

Not all subsets of an encoding combinatoric space can be computably implicitly defined in this differentiation sense by abstract machines: some will have characteristic functions that are highly uncomputable. The next critical step in this discussion, however, is to note 1) that the basic inability of any symbol string to represent a subset of possible symbol strings holds for any such subset that is unbounded, if not for simpler cases, and 2) that it is trivially possible for an automaton to implicitly define in this sense unbounded classes of possible input symbol strings. That is, it is trivially possible for a machine to implicitly define classes of input strings, classes of possible encoding strings, that cannot be represented by any string in the combinatoric space of such strings. Such subsets are outside of the limits of the representational power of any encoding combinatoric space. Such subsets can be *implicitly* represented, but not *explicitly* represented.

This is a general formal comparison between explicit encoding representation and implicit definitional representation of the sort that interactive representation is based on. Implicit definition is easily competent to implicitly represent unbounded classes, including unbounded classes of strings of presumed encodings, while explicit encoding representation is not. Encoding representation is limited to finite strings of whatever the atomic encoding set is taken to be.

Uncomputable Recognizers. Note that even for sets with uncomputable characteristic functions, a machine that would compute that function given unbounded time could still implicitly define the set for the purposes of *reasoning about* the set, even if not, in general, for recognizing instances of the set (Bickhard, 1973, 1980a). That is, such a machine or procedure can itself be taken as an object of knowing interactions, including reasoning, about any instances that satisfied that procedure — about any elements of the implicitly defined class — even if

the characteristic function is badly uncomputable. Mathematics and logic, among other things, are built on such abstractions by implicit definition (Bickhard, 1980a, 1988a, 1991d, 1992a; Campbell & Bickhard, 1986; MacLane, 1986); this is impossible within encodingism — encodings require *explicit* definitions.

Any *particular* implicitly defined set, however, *can* be given a derivative, stipulative, encoding once the crucial step of the implicit definition of that set has taken place. Note that the interactions of meta-knowing levels from which such implicit definitions can be explicitly considered — interactions of knowing system levels that take lower level knowing systems as environments of interaction — arguably constitute genuine reflexive knowing, reflexive consciousness (Bickhard, 1973, 1980a; Campbell & Bickhard, 1986), not the fake "reflexivity" that is really just recursivity of SOAR.

We argue that this greater power of implicit definition compared to explicit encoding manifests itself in a number of ways. If human representation is interactive representation, as the incoherence argument (among others) indicates that it must be, then attempting to capture human common sense reasoning and the representational powers upon which it is based within an encoding framework, such as that of Artificial Intelligence, will be tantamount to trying to capture the power of implicit representation with explicit encoding strings. It is no wonder that encodingism always encounters unbounded proliferations of new kinds of encoding elements (Bickhard, 1980b; Bickhard & Richie, 1983). The implicit cannot be captured by the explicit. This, incidentally, is another perspective on the impossibility of Lenat's CYC project: no set of atomic encodings can be adequate to the task.

Practical Implicitness: History and Context

To this point, the explication of the greater power of implicitness compared to explicitness has been at a formal level of abstraction. Of what relevance is it to actual problems? Are such unbounded classes of possible encoding strings ever of any practical importance? If not, then the inability of encodingism to represent such classes may be of no practical importance.

We argue that such classes of possible encodings *are* of practical importance in several senses, and that the encodingist ad-hoc and unbounded proliferation problem emerges in every one of them.

Collectively, we suggest, these manifestations of the proliferation problem include the frame problems.

Strings of encodings correspond, among other things, to *histories* of inputs to a system (again, note that interaction and timing are being ignored). Such histories, in turn, specify various contexts for a system. Differing histories may specify differing social or institutional situations that have been created in those histories — this is a real court proceeding versus a staged court proceeding for a film, or differing causal connections and relevances that have been constructed in those histories — the pressure switch attached to a bomb has been placed or has not been placed under my book, or differing intentional situations with respect to the intentions of other agents that have been affected or revealed in those histories — someone or no one is trying to kill me. Classes of such possible input histories, types of such possible input histories, correspond to types of possible contexts for the system. It should be already intuitively obvious that representation of such classes of possible histories, such types of contexts, is of critical importance. Again, this is within the power of implicit definition, but not of explicit encoding.

This formal transcendence of *context* representation, and, thus, context sensitivity representation, beyond the power of encodings is a formalization of the Shanon argument that context sensitivity precludes encodingism. The ubiquity of context sensitivity is a ubiquity of a kind of phenomena that cannot be captured within an encodingism.

Practical Implicitness: Differentiation and Apperception

Interactive differentiators are already implicit definers of potentially unbounded possible interactions. These classes of potential interactions may differentiate critical properties of the environment — properties that may not be differentiable by any finite class of inputs. Even a visual scan of a light pattern — for some solid object, for example — could in principle proceed in unbounded numbers of forms, with iterated rescans of various parts. If those rescans were to *not* yield the flow of visual interaction indicated by the original scan, then something might not be as it first seemed: the solid object that changed in some part is now seen to be not completely solid after all.

Similarly, the apperceptive procedures implicitly define the implicit situation image. The explicit situation image can never exhaust the implicit situation image, and certainly an atomized encoding shadow

of the explicit situation image can never come close to the implicit situation image.

Practical Implicitness: Apperceptive Context Sensitivities

A closer focus on this point will reveal, among other things, that apperceptive processes are themselves highly contextualized — and, therefore, highly historically context sensitive. Apperceptions are strongly externally contextualized: apperceptive processes are themselves heuristically driven by knowledge of what might be relevant in this situation, of what can be safely ignored or taken for granted in some other kind of situation, of what sort of apperception should return the most information in the current situation, and so on (Dennett, in press; Kyburg, in press). Apperceptive computation consumes resources, and the resource allocation process will itself be heuristically contextualized.

An example of this last point would be the situational dependency of apperceptions concerning a person laughing: if the person is seen walking alone down a sidewalk, the apperceptions will tend to be about the person, while if that person is seen coming out of a movie theater, the apperceptions will tend to be about the movie. A person diving to the ground on a city street will evoke quite different apperceptions than a person diving to the ground in a jungle in Vietnam in 1968. Interactive histories, and the contexts that they construct, guide the allocation of apperceptive resources toward what is heuristically indicated as providing the most important situation image constructions, toward where the most information is implicitly judged to be (Gilbert, 1989). In addition, apperceptions are also *internally* contextualized by interests, preferences, goals, values, appetites, and so on (Nutter, 1991). Within the interactive model, representation is *intrinsically* embedded in goals, which provides intrinsic internal contextualization for apperception: when a person is hungry, apperception will be more sensitive to possible indications of food.

Types of contexts, and their histories, can have quite complex relationships, including the relationship of some types constituting parts of, or historical points in, other types. Any establishment of a subconvention within a broader social convention would constitute one kind of example — the history (of utterances, usually) that creates a lecture versus a seminar convention within a broader class-meeting convention. Any setting up of one among alternatives of a subportion of a causal chain would constitute a different kind of example: Has the

bomb been directly wired to the pressure switch, or is there a radio link between them? Another sort of relationship is that of defeating relationships: The bomb is connected to the switch via a radio link, but there is a shield between the transmitter and the receiver.

In general, all knowledge is defeasible, including knowledge of what the defeasibility relationships are. Any causal connection or trajectory can in principle be interrupted; any social convention disrupted, faked, or overridden. It will always be possible in principle to cook up an exception or a conflict to any updating rule; Yale shooting problems are unbounded (Ford & Hayes, 1991). Unbounded classes of such contextualizing, and meta-contextualizing, of apperception can be implicitly defined in apperceptive procedures. It will be impossible to explicitly encode all such knowledge. Even such unbounded implicit definition, however, does not provide omniscience or prescience. Ultimately the apperception problem is one that is always subject to more learning, to variation and selection problem solving concerning successful contextualizations of successful apperceptions. Encodings are inadequate, then, in at least two senses: 1) explicit encodings cannot capture implicit knowledge, and 2) encodings cannot provide the error information for evoking learning processes.

A Counterargument: The Power of Logic

There is one possible counter to these claims that we would like to address. The basic claim of this counterargument is that encodingism has the power of axioms and inference, not just of encoding strings per se, and that the inferences generated by such a transduced-input encoding plus axioms plus inference, especially non-monotonic inference, can capture the unboundedness of representation being claimed for interactivism. There is a valid half of this point: axioms, axiom schema, and inference would in principle seem to capture the computational power of production rewrite rules, and this could capture the computational power of Turing machines. Neglecting issues of interactive outputs and timing, this would seem to give typical encoding systems the same power as interactive systems. In one critical sense, this is quite correct: a Turing machine, including one constructed out of production rules, can recognize — and thereby implicitly define — unbounded classes of input strings, just as in the case of other abstract machines.

Implicit versus Explicit. The difference is that such implicit definition is the heart of interactive representation, while encodings can

only represent that which they can derive or infer from the representational contents of inputs. Computation can "recognize" elements of, and, thus, implicitly define classes of strings, but that does not provide an *encoding* representation of that class — there is no way to specify *within the encoding space* what the representational content of the new encoding is supposed to be. To encode an unbounded class of possible strings on the basis of one finite member of that class is beyond the representational powers of encodingism.

Some special encoding atom, *not a member of the input alphabet*, could be invented just for the purpose of encoding that implicitly defined class. Then that special atom could be "inferred" on the basis of the recognition of a member of the class. But this is ad-hoc. It requires that such ad hoc encoding atoms be all presciently and mysteriously provided (innately?) or be designed in by the derivative observer/designer semantics. *And it is grounded, in any case, on implicit definition of that class of strings, not on explicit definition*. Still further, any such additional, new internal encoding atoms will simply generate a still larger internal combinatoric space, now not all strings of just *input* elements, *and the classes of strings of that larger space will not be encodable*.

That the special atom *is to be taken* as an encoding of that implicitly defined class of input strings cannot be defined within the representational power of the original combinatoric encoding space. It requires that implicit definition be applied to that space, and it requires that some observer/designer assign that implicitly defined content (of a class of strings) to that special encoding element. It requires that that special atom be given an *interpretation*. Implicit definition does not provide any representational content of what has been implicitly defined; therefore it cannot directly ground encodings — encodings require representational content of that which they are supposed to represent. Encodings can be based on implicit definition only via being grounded in something that can make use of representational implicit definition, such as an external human observer/designer. Or, the encodings could be derivative from a form of representation that does make use of implicit definition: interactive representation. Either option violates the assumptions of encodingism.

These points are just the axiom-and-inference version of machine recognition and implicit definition of the classes recognized. A machine — an automaton or Turing machine, for example — that recognizes, and thereby implicitly defines, a class of input strings does not thereby encode

that class; the machine has no representational content whatsoever about that class. Similarly, an axiom and inference scheme that recognizes and thereby implicitly defines a class of elements has no representational content about that class, and, therefore, cannot encode it.

Implicit definition is the core of interactive representation. Encodingism can capture the *computational* power of Turing machines via, for example, production rewrite rules, but encodingism cannot thereby capture the essential *representational* power without reliance on the utterly non-encoding property, the interactive representation property, of implicit definition. Within encodingism, computations can operate *on* representations, but computations cannot *be* representations. Even for capturing this computational power within an encoding framework, note that axiom *schema* are themselves a type of implicit definition of an unbounded class of axioms.

Grammars. An additional possible rejoinder to these points about the representational powers of encodings could be to claim, simply, that encodingism can represent unbounded classes of potential input strings, and, in fact, already has a well developed formalism for doing so: regular expressions. A regular expression is a means of representing unbounded iterations and embeddings of substrings within larger strings. "c*abc*" — for example — would represent "ab" flanked on each side by unspecified numbers of "c"s. This is an unbounded set of strings. Such regular expressions capture the classes of input strings that are potentially recognizable by a finite automaton. More powerful grammars, similarly, can characterize classes of strings recognizable by more powerful kinds of machines. In all such cases, the question arises: Why doesn't this count, why don't regular expressions and other grammars count, as representing unbounded classes of strings of input elements within an encodingism?

The short answer is: it does. But — all such grammars require their own dedicated elements, such as "*" for regular expressions, or "S" for a typical rewrite grammar, that **cannot** be elements of the input alphabet whose possible strings are being characterized. Insofar as such grammars constitute encoding systems, they are *meta*-systems with respect to the combinatoric space of the actual input elements. Their representational power, then, cannot be contained within any combinatoric space of any input alphabet, since they require special, *interpreted*, symbols of their own — interpreted with respect to

operations on, properties of, and relationships among, the strings and substrings in the input combinatoric space.

Encodingism, then, can representationally capture unbounded classes of strings of encodings *only* by transcending the representational power of *any* base of atomic encodings that it might begin with. This provides, in fact, still another version of the general point: given any set of representational atoms, define a regular expression (or some other grammar) on the strings of those atoms, and that constitutes a counterexample to any purported representational adequacy of that base of atomic encodings.

Furthermore, the move to a meta-encoding level within which the special notations for regular expressions and grammars can be defined for the base encodings is a move from atoms-as-that-which-is-to-be-represented to a language of encodings that represent them. This is a move that presupposes exactly all of the representational issues that are supposed to be at issue: How do those meta-level special elements acquire their representational content? The answer, clearly, is that they are defined and used and interpreted that way by human observers/designers/users. They constitute a *derivative* representational power — yet again — not an example of any power inherent in encodings per se.

Regular expressions and other grammars, then, can represent unbounded classes of strings of other elements, but only when they are provided the necessary interpretations by other intentional systems — humans. They constitute examples of both the power and the seductive danger of encodingism. The power is that, when a human being has created a new representational content, it can be assigned to some notational element, with consequent gains in clarity, manipulability, speed of processing, and so on. Such notational elements constitute encodings defined in terms of what they are to be taken to represent. The danger is in the confusion that such a process constitutes an example of encodings *creating* representational content, rather than of encodings being able to carry representational content once it has been already created elsewhere — by the inventor or user of the notation. The critique of encodingism does not question the ability, and the frequent desirability, of encodings to be invented for new representational contents — that, in fact, is exactly what is predicted. The critique does, however, demonstrate that such representational constructions, that transcend the representational power of *any* given set of atomic representational elements, can occur, do occur,

frequently occur, and that that construction itself *cannot* be explicated within encodingism itself.

Turing Machines. A related rejoinder would be to point out that a Turing machine can be encoded on the tape of a Universal Turing machine in the same alphabet as input strings — in 1s and 0s, in fact. This might seem to constitute the ability to encode the class of inputs that would be accepted or recognized by that Turing machine within the combinatoric space of the input alphabet in terms of such an encoding of the Turing machine itself. Clearly, the same point holds for finite state automata: the triples that define the transition function that constitutes the automaton could be defined in the input alphabet space.

The machine encodings, however, must themselves be interpreted as encodings of a machine — perhaps by a Universal Turing machine — and they must be distinguished from input strings per se — generally by position on the Turing machine tape relative to the read head at the start position. This provides the first perspective on the problems in this rejoinder: the machine encodings must be distinguished and interpreted *as* machine encodings and as *distinct* from normal input strings. The positional differentiation of a machine encoding from that same character string as an input string — as in the halting problem set up, for example — is, in effect, just a notational variant of the use of scope indicators and other operators, such as parentheses and stars, in grammars. The machine encoding cannot be just a string of input characters as input characters, else it would not be a machine encoding, and that distinction must somehow be made for the interpreting Turing machine. The machine encoding, then, may be identical in form to some string of input characters, but it cannot be permitted to *be* such a string of input characters if it is to be interpreted as defining a machine.

For an alternative perspective on this point, note that for the output of the Universal Turing Machine to be interpreted as the output that the machine-encoded-on-the-tape would give if it were run on the given input string — that is, for the standard interpretation of a Universal Turing Machine simulating the Turing machine indicated on the tape — the original string on the tape must itself be *interpreted* as consisting of an input string appended to a description or program or index of the to-be-simulated Turing machine. It is only with respect to the interpretations of certain strings as indexes or descriptions of Turing machines that the proof of and the very notion of Universal Turing Machine can exist at all. This interpretive step, and, therefore, anything based upon it — such as

the interpretation of the Turing machine index as being, in addition, an implicit definer of the class of strings that that Turing machine could recognize — is *not* within the purview or competence of the Universal Turing Machine itself. As usual, such interpretations, currently, are provided only by human beings. They cannot be captured in encodingism.

A second problem with this Turing-machine-index-on-a-tape rejoinder emerges even if this point about interpreting the index ***as*** a Turing machine is overlooked. The machine encoding does provide a correspondence to the implicitly defined class of input strings that would be accepted or recognized by that machine — just as a grammar provides a correspondence to the implicitly defined class of strings that it could generate. If the machine encoding were taken as representing the implicitly defined class of recognizable input strings, then that constitutes three different functional interpretations of the same string: 1) as input string, 2) as machine encoding, and 3) as class-of-input-strings representer. These interpretations must be kept distinct for the *interpreting* machine, and, in particular — as above — the definer of a *class* of input strings cannot be permitted to be an input string per se (in the same combinatoric space) itself. So it is still impossible to *represent* a class of input strings in general within the combinatoric space of the *input encoding alphabet*.

The third and deepest problem with this rejoinder has already been alluded to: taking a machine-defining encoding as representing the class of strings that that machine could accept or recognize is, in principle, no particular problem for humans, but to do so is precisely to make use of implicit definition, not explicit definition. Any such usage transcends the boundaries of what can be defined with the input encodings as encodings — the *formal character* equivalence with an input string does not give the machine encoding, thus the implicit input string class definer, any possibility of being *representationally* defined, defined as an encoding, within the combinatoric space of input strings. Such a possibility of notational equivalence between semantically non-intertranslatable sentences is already well known. Gödel's theorems, for example, depend on it.

In general, then, the fact that a machine can be encoded in the same alphabet as that machine's inputs does not provide a way to represent the class of acceptable strings to that machine or for that machine or within the input encoding space. The machine encoding

might be in the formal input *character-string* space, but will not and cannot be in the input *encoding* space. And, even if a machine encoding *is* taken as representing the implicitly defined class of acceptable strings, that interpretation requires implicit definition, which is outside the boundaries of possibility within encodingism.

Concerning representational power, then, encodingism is in a dilemma. Either it is not competent to the unbounded implicit representations of interactive differentiations, apperceptions, and indications, or else it makes use of implicit definition in its own axiom schema and computations and interpretations, and thereby violates encodingism altogether. An encoding must carry known representational content; implicit definition does not provide that, and, therefore, cannot directly ground encodings. Implicit definition can ground interactive representation, however, because interactive content is not given in what is implicitly defined per se, but rather in the *indications between* the implicit definitions of differentiators and the implicit definitions of potential further interactions. An interactive indicator implicitly predicates that any environment in the implicitly defined differentiation class will also be an environment in the implicitly defined further-interaction class.

Incoherence: Still another corollary

The proliferation problems of encodingism provide a perspective on still another corollary of the incoherence problem. Encodingism is an approach to representation from within a broad metaphysics. Encodings represent explicitly; they represent finitely; they represent actualities; they represent objects, events, and so on. They presuppose a substance ontology, whether atomized or not. Encoding atoms are intrinsically static — Wittgenstein's atoms in the Tractatus were necessarily unchanging: if they could cease to be, and if the encodings' meanings *were* their correspondences to these atoms, then the existence of the meanings of the encodings would depend upon matters of fact of whether or not the corresponded-to atoms in the world still existed. Encodings bear representational content as primary properties.

Interactive representation, on the other hand, represents implicitly, potentially unboundedly. Interactive representation is of potentialities, not actualities — interactive potentialities, in fact. Note that interactive representation cannot be caused by that which it represents: what is represented is potentiality, not actuality, and potentiality does not yet

exist to be able to yield such causal consequences. Interactive representation is embedded in a process ontology, and, as such, is intrinsically dynamic. Interactive representational content is not a primary property of an atom or substance, but an intrinsically relational property, a functional property. A pure substance and property metaphysics cannot be competent to such relational issues (Olson, 1987). Encodingism and interactivism are on opposite sides of a quite vast divide.

These metaphysical characteristics of interactivism are not an accidental collection of properties; they are intrinsically and necessarily related. Interactive representation is a functional property of interactive dynamic systems, a property of particular organizations of system process. It cannot be of actualities because it is of process — of action, of interaction — and those are potentialities: at best the current point in a current interaction is "actual," but such "points" are not what is interactively represented. Interactivism, in intrinsically representing potentialities, is intrinsically modal. Potentialities are potentially unbounded, therefore interactive representation cannot in general be exhaustively explicit. Encodingism fails, and necessarily fails, on every one of these points.

This metaphysical perspective provides another aspect of, and thus another corollary to, the incoherence problem. Encodingism requires that its atoms bear their representational contents independently. Yet, if interactivism is correct, representational content is constituted in *functional relationships*, functional indications in system organization. Encodingism, then, requires that its atoms bear their own functional relationships independently of what they are relationships to; they must provide their own functional potentialities. But functional potentialities are relative to the functioning of particular systems. Functional potentiality, thus representational content, cannot be defined independently, atomistically, because it is an intrinsically relational property. Encodingism involves the incoherence of independent atoms bearing intrinsic functional relationships. It requires that its atoms bear functional relationships that dangle in a logical void, not being related to anything.

Counterfactual Frame Problems

The point that modality generates unboundedness which generates frame problems holds as much for potentialities that are *not* anchored in,

or are interactively *not* accessible from, the current situation as it does for those that *are* potentialities of the current situation. That is, it holds for counterfactual considerations as much as for accessible potentialities (Stein, 1991). If kangaroos didn't have tails, they would fall over — unless they were skilled with crutches; if my computer were an IBM 7094, it wouldn't fit in my room — unless I had a much bigger room. Counterfactual frame problem proliferation and defeasibility issues are quite similar to cases that do tie together with the current situation, such as for possible actions.

Such counterfactual reasoning involves classes of possible worlds — types of possibilities — and the relationships among them. Encodingism provides a very uneasy and ultimately inadequate approach to such issues. It makes no sense to assume that each possible world is distinctly encoded, for example, and the accessibility relationships among such worlds — what can be assumed to be *the same* as in this world in a world in which kangaroos have no tails? — are just as atomized when rendered in an encoding approach as are the interactive relevancies in the situation image. Similarly, those encodings of accessibility relations between worlds — in object based models, something like Transworld Heir Lines (Kaplan, 1979c) — must be built back in in just as ad hoc a manner, and with the same sorts of proliferation problems as in the original frame problem case.

Interactivism is intrinsically embedded in modality, and provides a natural approach to issues of modality, of types of possible worlds (Bickhard & Campbell, 1992). An interactive implicit definition **is** an implicit definition of a class of possible situations, of a type of possibility. They constitute differentiations within the space of possibility. They are of types and forms, not of singular bare particulars — there are no bare particulars (Loux, 1970). There will be hierarchies of such situation image *scheme types* (Bickhard, 1980b), with a natural equivalent of inheritance of interactive properties down the hierarchies. Intersections of the differentiation hierarchies provide refinements of the differentiations of the world, and provide a topology on the types of possibilities represented. This topology provides a natural approach to the issues of what constitutes a "nearby" class of possible worlds. One encoding atom is just as near to or far from a given atom as any other — there is no natural topology among encoding atoms — but the intersections and overlaps of *interactive* differentiations provide *neighborhoods* of the current situation that constrain such issues of

similarity of possibilities. Possibilities in smaller neighborhoods are nearer than possibilities in larger such neighborhoods. Note that, in principle, this neighborhood point holds for general locations in the space of possibilities; it is not specific to the location of the current situation.

The technical issues involved here can clearly become complex, but the basic point is that the intrinsic involvement of modality in interactive representation not only explains the proliferations of the basic frame problems, but it also thereby provides powerful resources for counterfactuals and modalities in general, and the frame problem variants that they can generate. The implicitness, unboundedness, dynamic nature, and modality of interactive representation are all intrinsically related, and provide an approach to modality in general. The actual current situation is "just" a location within that broader organization of potentiality of interaction.

The Intra-object Frame Problem

Apperception is not only of changes in the properties and statuses of objects, it is also of the objects themselves. This yields an intra-object frame problem, not just an interobject frame problem, and pursuing that frame problem yields still another argument for interactivism.

If I move this book, that wall will still remain — generally. But also, if I move this book, its pages and its back cover will remain. In fact, if I hide this book under a cover, its front and back and pages and so on will still remain invariant — generally. These points are just as much apperceptive as those between the book and other objects. They constitute knowledge that takes a couple of years in infancy to develop (Piaget, 1954).

But this has serious consequences for encodingism. If intra-object apperception generates its own potential frame problems, then it is simply a papering over of ignorance to pretend that encodings correspond to and thus represent objects (and their properties and events, etc.). On the other hand, if we try to find something of an intra-object character that we can assume such correspondences with, it cannot be found. We are on a search for early Wittgenstein's Tractarian atomic objects. Any causal connection can in principle be blocked, any causal potentiality defeated. Any assumption about such a presumed metaphysical atom for encoding correspondences to latch on to will involve presumptions about the potentialities of these atoms for further interactions, including further sensory interactions. Any of these potentialities could be potentially

defeated, and the encoding of the conditions under which they are to be taken as not defeated and those under which they are defeated generates an intra-object frame problem. Any presumed encoding correspondence with *anything* necessarily involves presumed potentialities of action, of interaction, of perception, of apperception concerning what those correspondences are with. And that necessary irruption of potentiality, of modality, into representation destroys such presumed correspondences — destroys them with frame problems concerning the apperception of what the correspondences are supposed to be with.

Ultimately, there is nothing left for encoding atoms to correspond to. There are only points in organizations of potential further interactions, and the apperceptions of those organizations in situation images. Actuality does not *possess* potentiality; actuality is *a location within the organization* of potentiality — the organization of potential further interactions. Therefore, actuality — the actual current environment — cannot be represented independently of potentiality, and, therefore, cannot be directly encoded. Modality, potentiality, generates the unboundedness that generates the frame problems; it can be captured only with implicitness. Therefore, it can be captured only within interactivism.

11

Language

Throughout history, theories of language have focused on two fundamental issues: 1) the relationship between language and thought, and 2) the nature or purpose of language. In this chapter, we examine how Artificial Intelligence models of language have addressed these issues. We also review and elaborate our sketch of the interactivist model of language to inform our critique of AI models.

Theories of language and theories of thought have tended to cluster. Language often has been taken to have some sort of privileged relationship to thought. This intimate relationship has meant that how one views language helps shape how one views thought, and vice versa. As we have discussed at length, the classic, widely accepted model of thought has been *encodingism* — thought consists of processes operating on encoded structures. Encodingist models of thought have led to *transmission* models of language — language consists of a speaker re-encoding mental structures into structures suitable for linguistic transmission to a hearer, who then must decode them (Bickhard 1980b; Winograd & Flores 1986). Encodingism and transmission theories have engaged in an interesting feedback loop; Jerry Fodor (1975) has systematized encodingist assumptions about thought by referring to the "language of thought." In transmission models of language, language consists of formal productions of formal *public* structures; thought becomes formal productions of formal — but *private* — structures.

However, there have been alternatives to the transmission view. These have tended to focus on the social, functional, and ontological aspects of language. Whorf (1956), for example, argued that the vocabulary and structure of a language influenced the thought and actions of users of that language. Vygotsky (1962, see also Wertsch, 1985) focused on language as a tool, first, for socializing a developing child into a community, then as a means for regulating one's own thought.

Two other approaches are of more relevance to our discussion: Heideggerian philosophy and ethnomethodology. Heidegger and his followers see language as social action that creates webs of commitment between people. They argue that reality is (at least) largely constituted through language: what *is* is what we can talk about. We discuss this approach in more detail later in this chapter when we consider the work of Terry Winograd and Fernando Flores, who have developed a Heideggerian critique of Artificial Intelligence and an alternative research programme for computer science. Ethnomethodological approaches to language are convergent with Heideggerian views in their focus on language as social action; in addition, they are strongly empirical, observing and describing the phenomena of actual language use. This has resulted in important insights into the dynamic organization of activity, including language activity. These insights have begun to be applied by computer scientists seeking a new foundation for intelligent systems (Luff, Gilbert, & Frohlich, 1990). We consider this work in detail in our discussion of Lucy Suchman.

Another set of fundamental issues concern the nature or function of language. Encodingist models of language have seen language as a *cognitive* phenomenon, operating *within a single individual*. They have seen the purpose of utterances as being to *convey information*, thus the focus on *truth conditions* of sentences. Finally, they have idealized the meaning of language as being independent of context, taking proper names as paradigmatic cases.

Alternative models have seen language primarily as a *social* tool, inherently requiring reference to *all participants in a conversation* for accurate description and understanding. They have focused on *language as action*, as designed to achieve some purpose. They have pointed out the many ways in which language depends on context, taking indexical speech as paradigmatic.

Bickhard (1980b, p.14) categorizes the relationships between language and thought as shown in the following matrix:

	Transmission	**Transformation**
Encodingism	naturally compatible	forced compatibility possible
Interactivism	incompatible	naturally compatible

We do not reproduce here the full logic of the argument underlying this matrix (see the discussion above; Bickhard, 1980b, 1987). We note

however, that the table suggests an important dynamic: as models of language become more transformational, i.e., focus more on the social, operative (action-oriented) and context-dependent aspects of language, they match less well with encodingist models of thought and knowledge. They make such models seem inadequate and exert pressure to modify the models.

This dynamic can be observed in AI. Models of language have become radically more transformational over the last two decades. However, while the natural affinity for transformational models of language is interactive models of thought, it is difficult to make the leap from encodingism to interactivism. Encodingism is usually implicit and almost always deeply presupposed within Artificial Intelligence and Cognitive Science. Therefore, while we consider AI models of language to have evolved in directions compatible with interactivism, and while this has exerted pressure on AI models of language that have moved them in the right direction, they remain essentially encodingist. It takes a *revolution*, not an *evolution* to move from encodingism to interactivism.

INTERACTIVIST VIEW OF COMMUNICATION

The interactivist model of thought precludes a transmission model of language: Because knowledge does not consist of encoded structures, language cannot consist of the re-encoding, transmission, and decoding of these non-existent structures. Instead, interactivism sees language as a "social resource for the creation, maintenance, and transformation of social realities." (Bickhard, 1987).

The heart of the argument that language operates on social realities goes as follows. We might begin by assuming that language operates on other minds. However, if the *direct* object of an utterance were the mind of the hearer (or audience), then the successful completion of an utterance would be dependent on the effect it had on the mind of the hearer — a command would not be a command unless it were obeyed, nor an assertion an assertion unless it were believed. Instead of making mind the direct object of language, interactivism proposes the construct of *situation conventions* (Bickhard, 1980b, 1987, 1992a) — intuitively, socially consensual definitions of the situation. Thus, an assertion changes the definition of the situation. For example, it can allow the hearer to believe that the speaker believed what he or she said, it may cause the hearer to attempt to determine presuppositions that are necessary in order for the assertion to make sense, and it commits the

speaker to support or explain the assertion if the hearer is unsure why the assertion makes sense or disagrees with it.

The need for language arises from people's need to *coordinate* their definitions of the situation. Such a condition of coordination is called a *situation convention*. Recall that in interactivism, one's knowledge is constituted by the range of potential interactions. And in the presence of another person, the range of potential interactions is constrained and constituted by that person, in particular, by that person's definition of the situation. Thus, two people need to coordinate their individual definitions of the situation (both implicit and explicit) in order to manage the space of possible interactions, and this is the fundamental function of language. When such coordination is achieved, it constitutes a situation convention.

To summarize, language is inherently (ontologically) *operative* and *social*. We now proceed to explore several consequences of the basic model.

Two important implications derive from the fact that utterances are operators. First, the meaning of utterances are inherently context-dependent, since, in general, the result of an operator depends on the operand(s) it is applied to. This offers a natural way to explain why saying "Can you pass the salt?" or "It's cold in here" can be interpreted differently in different circumstances. Second, the meaning of an utterance *type* is taken to be its operative power, rather than the result of an utterance of an *instance* of that type. As an operator, an utterance token does not have a truth value, but the *result* of an utterance — a situation convention — can have a truth value, since it represents a situation and can do so more or less correctly.

Third, the fact that utterances operate on situation conventions, together with people's need to coordinate their situation conventions, offer a way of accounting for phenomena like presupposition and implicature (Bickhard, 1980b; Grice, 1975; Levinson, 1983). Often, in order to make sense of an utterance, in order to determine how the utterance *could be applied* in this situation, the hearer is forced to adjust his or her definition of what the situation is. In interpreting what presuppositions could have been involved in order to make an utterance appropriate to a situation, the hearer may come to share further presumptive commonalities about that situation with the speaker — in such a case, the utterance will operate on the situation convention via implicature.

Fourth, the interactivist model makes it clear that the primary object of interaction is the situation convention, a social object. Therefore, the ontology of the person is largely social and, because the social is largely linguistic, the ontology of the person is massively linguistic (Bickhard, 1992a; Campbell & Bickhard, 1986, p. 127). Again, knowledge consists of indications of potential interactions, and since the single most important object of interaction is other people, our knowledge is largely social. And since language is the tool by which and the medium through which social relationships are constructed, expressed, and experienced, language is at the heart of what we are.

This analysis has deep convergences with Heideggerian and hermeneutic philosophy (Heidegger, 1962; Howard, 1982). There also is one important distinction. Hermeneutics often lapses into a form of social/linguistic solipsism, in which nothing exists outside the bounds of language communities. Interactivism, however, provides means of grounding the ontology of the person prior to linguistic communities. As Bickhard (1987, p.45) puts it: "Being is that which codetermines the outcome of our interactions." In chapter 7, we make it clear how interactivism provides a grounding for knowledge in the world. When we discuss the work of Winograd and Flores later in this chapter, we elaborate on the distinctions between hermeneutics and interactivism in this crucial respect.

Finally, we note that linguistic interactions are a special case of the family of general goal-directed interactions, and, as such, they inherit certain properties. One such property is that, in the limiting case, blind trial-and-error variation and selection must play a role. We do not know a priori how to express or interpret everything that can be said. Difficult language such as highly ambiguous texts, historical writings, or psychotherapeutic conversations illustrate this. Therefore, models of language that are strictly algorithmic or presuppose fixed meanings for words or other linguistic constructs are inherently inadequate.

THEMES EMERGING FROM AI RESEARCH IN LANGUAGE

The discussion of AI language work will focus on a number of ideas of which AI language researchers have become aware over the last 25 years. These ideas lead away from a strictly encodingist, transmission view of language toward a more interactive, transformational conception.

Awareness of the Context-dependency of Language

By the mid 1970s, AI researchers had developed a strong focus on studying natural types of language use, for example, translation or story understanding. This was a major step away from the methodology of philosophers and linguists, who usually studied contrived single-sentence examples. This led them to the realization that knowledge of both the physical and social worlds and of the conventions of language interaction was crucial in building systems that could use language in human-like ways (see, for example, Bobrow et al, 1977; Minsky, 1981; Schank & Abelson, 1977; Winograd, 1976; and Waltz, 1982 for a summary article). An example from Winograd (1972),

The city councilmen refused the women a permit because
(a) *they* feared violence.
(b) *they* advocated revolution.

showed the necessity of a great deal of knowledge about the social world in order to interpret what the pronoun "they" referred to. Such realizations helped to make the issue of the organization and use of knowledge the central topic in AI research. Language processing was the first area within Artificial Intelligence that discovered the need for large quantities of real-world knowledge in intelligent activity.

Awareness of the Relational Distributivity of Meaning

Researchers began to devise knowledge structures to capture the properties of human knowledge that were known and to support the process of language understanding. One of the first such structures was semantic networks (Findler, 1979; Quillian, 1968; Simmons, 1973), which formalized knowledge in terms of nodes representing concepts and arcs representing relations between those concepts. Semantic networks and the processing technique of spreading activation were designed to capture intuitions of associations between concepts, and how thinking about one concept could activate "close" — related — concepts.

Frames (Minsky, 1981; Schank & Abelson, 1977) were a related notion that began from the premise that knowledge was organized into situationally related chunks, or models of aspects of the external world. For example, there would be frames for rooms, for sequences of events in a restaurant, for aspects of airline trips, and for every other chunk of the world that someone knows about. Frames are defined in terms of an organization of slots and fillers. Slots represent aspects of a class of situations that can vary from instance to instance, like the destination in

an airline trip. Fillers represent the values for a particular slot in a particular situation, e.g., the destination of an airline trip might be San Diego. Processing information involves assembling a configuration of frames that offer the best account for the input. The set of instantiated frames constituted the interpretation. Below we discuss GUS (Bobrow et al., 1977), a well-known early attempt to use frames to construct a language processing system.

Frames captured many important intuitions. For example, a slot could restrict the type of fillers it could take, allowing the system to reason about whether a given value could plausibly appear there. Also, a slot could specify a default value. These capabilities meant that when some input activated a frame, other parts of the frame could also be added to the interpretation. For example, seeing just the handset of a telephone could trigger a phone frame, which would add the information that there was a body of the phone, a cord, etc. In addition, as discussed below, various types of procedures could be associated with the slots of a frame. All in all, this led to a much more active conception of knowledge.

In general, experience with more complex and active knowledge structures caused the Artificial Intelligence idea of meaning to become much more complicated. The main insight of this period was that the meaning of a symbol inhered in two things: its relations to other symbols (defined by a path along the arcs or slots of the knowledge representation), and the set of procedures that operated on the knowledge structures, which defined how the knowledge structures could be traversed and combined. Thus, Artificial Intelligence was developing a relationally distributed and procedural view of knowledge. How these knowledge structures were to be grounded in the world, however, remained a mystery. By this point in the discussion, the reason for this mystery should be apparent: the encodingist conception of representation fundamental to AI cannot provide a representational ground for internal data structures.

The PDP approach of the 1980s has pushed this distributed view much further, successfully accounting for many aspects of knowledge that frame and semantic network theory aimed at, but which symbolic implementations were unable to attain. However, as discussed below, the underlying semantics associated with PDP networks is the same as that for conventional symbolic Artificial Intelligence systems.

Awareness of Process in Meaning

Active knowledge — procedural attachment. Procedural attachment refers to the association of procedures with the slots of a frame. Two major types of procedures are **if-needed** procedures, which constitute local heuristics for finding a plausible value for a slot, and **if-added** procedures, which ripple a series of effects through the knowledge base when a value is added to a slot. Thus, associated with the *political-Party* slot of a person there might be an if-needed procedure embodying the heuristic "if they live in a rich neighborhood, or drive an expensive car, or are employed as a banker, then assume that they are Republicans." In turn, adding a value to the slot *political-Party* could cause other values to be added elsewhere. In general, procedural attachment maintains coherency and consistency across a knowledge structure, embodying common relations between pieces of knowledge, e.g., co-variation of political sympathies with wealth or type of car owned. If-needed procedures also bring an element of goal-direction to the interpretation process. Finally, procedural attachment makes knowledge more active and process-oriented. That is, it is not sufficient to simply capture the structure of some phenomenon: one must also define procedural relationships and triggering conditions for deriving these relationships.

Procedural Semantics — Language. Procedural Semantics was born with an analogy between natural languages and computer languages. Every first year programming student is taught that the language she is learning has both syntax and semantics. The syntax is a set of rules specifying how one can arrange the semi-colons, parentheses, and assorted other symbols to form legal programs. The semantics is what happens when the program is run. There are two phases to running a program written in a high level language such as Pascal or Lisp. One is compiling the program, that is, translating it into a form that actually can be run by the machine, and the other is running it.

The analogy then goes like this: in some sense, we all agree that natural languages have both a syntax, which we understand pretty well, and a semantics (here, we mean something neutral enough to please everyone; "meaning" would be a more theory neutral term), which however, we don't understand well at all. But if we turn to programming languages for inspiration, we could say that, just as a statement in a programming language goes through two stages to be understood by the machine, so too must a statement in a natural language go through two similar phases. The first phase is to go from the sentence in the natural

language to some procedure *in the internal mental language*. The second is to execute this procedure (with the caveat that the hearer gets to decide whether to execute it). And, indeed, for many sentences, particularly questions and commands, the analogy seems to work quite well. For example, "Can you pass the salt?" might go over into a procedure that, when executed, causes the hearer to pass the salt.[14]

Procedural semantics was logically strongly committed to the context dependency of language, in much the same way as interactivism (although we know of no one who has made this particular argument). Since an utterance is a program, and a program produces different output from different inputs, an utterance must have different effects when uttered in different situations. However, while interactivism uses the object of language, situation conventions, as a powerful constraint on its theory of language, until recently Artificial Intelligence work had not addressed directly just what the inputs and outputs of "utterances-as-programs" should be.

In addition, just as interactivism sees the process of understanding an utterance as consisting of a transformation of the internal state of a system, so too must executing a program result in some change of state. In Artificial Intelligence perspectives, however, that state might be something as obviously encodingist as a list of propositions, and is *always* some sort of structure of encodings. Thus, while interactivism gives the changed internal state representational content by its role in indicating and constraining future interaction, in Artificial Intelligence models the representational content of internal symbols was merely assumed.

This leads to the somewhat paradoxical observation that procedural semantics is not semantics, at least not in the classical Frege-Russell-Carnap sense of the term. As Fodor (1978) pointed out, while classical semantic theorists at least attempt to explain the *aboutness* of language, how (for example) we can use a symbol like "dog" to refer to a dog, procedural semantics is silent on this issue. Indeed, as remarked by Johnson-Laird in later work (1983), Procedural Semantics does not relate language to the world but rather to a "mental model" or internal representation. This is the view that is dominant in AI language work.

---

[14] A clear statement of these ideas is found in Johnson-Laird (1977), which sparked a critique by Fodor (1978), a reply by Johnson-Laird (1978), and a commentary by Wilks (1982). Winograd's SHRDLU program (1972) is the best known and one of the earliest examples of treating utterances as programs.

Such a view of language processing as utterances invoking functions that manipulate mental models is partially consistent with interactivism; however, *AI and interactivism differ crucially in their explication of "mental models"* — this is the crux of the distinction between encodingist and interactivist notions of representation. Artificial Intelligence work on language takes as given the existence of symbols such as MOVE, CAUSE, PART-OF, and explores the sorts of reasoning that can be done with them.[15] It is commonly said (for example, Wilks, 1982) that the meaning of such an item is all the inferences licensed by it. This idea of "meaning as use" is compatible in broad outline both with Wittgenstein in *Philosophical Investigations* (1958) and with interactivism. However, interactivism takes the crucial additional step of providing a grounding for the entire system in the world via an interactive model of representation. So, Artificial Intelligence and interactivism partially share conceptions of language use, but interactivism provides an account of how the representations that are the objects of language are grounded in the world. AI provides, and can provide, no such account. Thus, at best AI offers a simulation account of representation.

An additional caveat concerning the similarity between the AI and the interactivist conceptions of language use derives from the fact that AI considers the processes that transform from utterance to internal program to be more or less fixed and sequential in nature while this is, in general, impossible from the interactivist perspective. Certainly much quotidian language has the practiced and habituated character that gives it an algorithmic flavor, but any difficult language — ambiguous text, historical text, psychotherapy, language learning, and so on — reveals the underlying variation and selection constructivist character of language understanding.[16] Our suggestion, in fact, is that this trial and error constructivist character of language understanding is what is referred to as the hermeneutic circle (Gadamer, 1975; Heidegger, 1962; Howard, 1982; Ricoeur, 1977).

Such a foundational character of language understanding is necessary from the interactivist perspective since language understanding — interpretation — is a special case of general apperceptive processes,

---

[15] David Waltz (1982) said "such systems cannot be said to know what they are talking about, but can only know *how* to talk about things."

[16] In the interactivist analysis, *all* knowledge must, in the limiting case, be based on blind variation and selection — to assume otherwise is to attribute prescience to the agent or system being analyzed. See the discussion on interactivism and variation and selection in Bickhard (1992a).

and apperception is necessarily of variation and selection constructivist character because the system cannot in general have prescient foreknowledge of which perceptual and apperceptual processes are appropriate, so it must try some out in order to determine what fits the perceptual selection pressures of the current environment (Bickhard, 1992a). In many cases, of course, some degree of such foreknowledge *will* be present based on earlier interactions, and it is by focusing on these cases that the algorithmic view retains appeal.

Case study: GUS. GUS (Bobrow et al., 1977) was a project that explored issues in language processing using frames as a representational technology. GUS simulated a travel agent assisting a client in choosing a flight. GUS addressed the problem of structuring knowledge with frames, which it used explicitly to represent world knowledge and implicitly to structure the dialogue. For example, a frame for an airline trip might look like:

TripSpecification

homePort	*isa*	**City**
foreignPort	*isa*	**City**
outwardLeg	*isa*	**TripLeg**
awayStay	*isa*	**PlaceStay**
inWardLeg	*isa*	**TripLeg**

The structure of the frames provided a guide to controlling a user-system dialogue: a place had been prepared for each piece of information relevant to booking a flight, i.e., there was some slot of some frame in which to put it. Thus, a simple dialogue control regime was for GUS to try to fill in its slots in the order in which they occurred.[17] In general, the set of frames and slots possessed by GUS (or any GUS-like system) defines its goal space, where its ("hard-wired") goal is taken to be "fill my slots." In addition, GUS could handle information volunteered by the client. This control regime permitted quite realistic mixed-initiative dialogue to occur.

GUS illustrates the power of frames-based processing in simulating aspects of language use. It does so in ways that are partially convergent with interactivist ideas. First, GUS explicitly maintains a conversational context in the form of the travel frame that GUS has

[17] Technically, the slots were filled depth first. That is, each slot could itself be filled by a frame. So if for example, frame F consisted of slots S_1,..., S_n , and S_i's filler had to be a frame of type F_i, then GUS would try to fill slots S,...,S_{i-1} then would fill all the slots of F_i, which itself might involve recursion, then would return to filling F's remaining slots.

managed to construct at any particular point in the dialogue. Second, utterances are operators on this context or "frame change descriptions." For example, the utterance "I want to go to San Diego" goes into the frame change description (informally) as:

Add a frame F1 which *isa*
TripLeg, whose
Traveler is the client,
Destination is a frame F2 which *isa* **City**, whose
Name is "San Diego", and whose
TravelDate is a frame F3 which *isa*
Date, whose
Month is "May", and
Day is 28.

And third, the structure and possible contents of GUS's frames plus its dialogue control regime defined the simple social interaction of a ticket agent helping someone to arrange an airplane trip.

However, the frame-based approach that gave GUS its power has fundamental limits. First, at best, it illustrates a *transformation of encodings* approach to language. While this is a logically possible approach, as we argued above, it is severely limited in that it inherits all the (insurmountable) problems of encodingism. A second and more crucial point derives from the first: GUS's knowledge was limited to the frames and procedures that its designers had supplied it with. It would be unable to respond to questions outside this space, such as "Would it better to take a train or a plane from New York to Boston?" Of course, the designers could supply it with more frames and procedures to handle such questions, but the resulting system would still have strictly limited, albeit expanded capability.

A standard AI rejoinder to the problem of the limited competence of any frame-based system[18] would be: give it more frames! Give it frames about *everything* that could conceivably be relevant! This is what Doug Lenat has tried to do in the CYC project (and we discussed above why this is no solution). Terry Winograd was one of the researchers involved in the GUS project. His reflections on the shortcoming of frame-based systems in particular and Artificial Intelligence in general

[18] We argue, of course that these problems are inherent to encodingist systems of all sorts, whatever the details of the encoding structures and operators.

eventually led him to a radical break with AI and turn toward Heideggerian philosophy.

GUS is important as a specific attempt to build a system that embodied some of the assumptions about the importance of context and procedures in language understanding. However, in addition, as part of the project, observations were made of people interacting with a person simulating GUS to see what sorts of phenomena occur in natural dialogues of this sort. These revealed several shortcomings.

This and other studies like it (Tennant, 1980), began a journey from concentrating on language understanding systems to working on natural language **interfaces**, systems that interact with people. In the early '80s, researchers in Computer Science, Human Factors, Cognitive Science, and the social sciences joined to follow this idea, leading to the emergence of Human Computer Interaction as the study of how to design computer systems that people can interact with more easily. This in turn has led to a great deal of interest in and studies of just what makes natural human interaction so effective. This work has resulted in deep critiques of Artificial Intelligence models of action and language and has been used by researchers interested in constructing a new foundation for intelligent systems (Fischer, 1990; Hollan et al., 1991; Lai, Malone, & Yu, 1988; Stefik, 1986; Terveen, 1993). Later in this chapter, we illustrate this research using the work of Lucy Suchman as an example.

Toward a Goal-directed, Social Conception of Language

Artificial Intelligence models of language have long had an action-oriented flavor. In order to construct a practical system that could respond sensibly to natural language requests about a database of train schedule information (for example), issues such as understanding the intent of a request and knowing what information could usefully serve that intent were crucial. The truth conditional analyses beloved of philosophers were of little use in such a project. However, speech act theory (Austin, 1962; Searle, 1969), which focused on the purposes of language, did offer useful guidance. In the next discussion we focus on how Artificial Intelligence researchers have formalized speech act theory in terms of AI *planning* and used this as the basis for models of dialogue.

While an action-oriented perspective has become mainstream in Artificial Intelligence language studies, the perspective has still been largely *cognitive*, focusing on the internal processes used by a speaker to produce language or a hearer to interpret it. While much work is

premised on the notion of *user models*, we will argue that this is a pale, encodingist shadow of the actual social and situated nature of language.

Awareness of Goal-directedness of Language

Since the mid 1970s, much Artificial Intelligence work has built on the foundation of speech act theory. The fundamental recommendation of speech act theory, that language be seen as action, also underlies the interactivist account of language. In addition, some developments of speech act theory, notably by hermeneuticists (Habermas, 1971, 1979; Winograd & Flores, 1986), are particularly compatible with the interactivist approach; for example, that speech acts occur as part of a network of commitments to action by participants in some conversation and that language is a tool that agents use to coordinate their actions.

Major themes of action-oriented approaches within Artificial Intelligence have been that utterances are produced in service of an agent's goal and that responding appropriately to utterances requires inferring from them the underlying intentions of the speaker. For example, Allen (1983) studied cooperative behavior by a train clerk, who, when asked "When does the train to Montreal leave?," replied "3:00 at gate 7," volunteering the location of the train since he had reasoned that the customer must have the intention of catching that train, and that in order to realize this intention, the customer needed to know the location. Another typical concern has been to compute appropriate responses to indirect speech acts. Suppose a customer asks our clerk "Do you know when the train to Montreal leaves?" Then the appropriate response most likely is "Yes — at 3:00, gate 7."

AI Planning. Planning has received an enormous amount of attention in the AI literature, beginning with Newell and Simon's GPS, progressing on through various formalisms and algorithms such as STRIPS (Fikes & Nilsson, 1971), WARPLAN (Warren, 1974), INTERPLAN (Tate, 1974), NOAH (Sacerdoti, 1977), and TWEAK (Chapman, 1987). A planning system consists of a language for describing "states of the world" or situations and a library of operators, each of which transforms one state of the world into another. Operators are defined in terms of (at least) *prerequisites*, *decompositions*, *effects*, and *constraints.* Preconditions are conditions that must hold of the world (or be made to hold) before the operator can be applied. If the preconditions are not true, the planner may try to make them true.

Constraints, too, are conditions that must hold of the world before an operator can be applied. Unlike preconditions, though, if a constraint does not hold, the planner does not try to achieve it. Effects are conditions that will hold after the operator has been applied. Decompositions represent more primitive operators, which when performed together, constitute the performance of a single operator. For a simple example, **move** might be axiomatized as:

MOVE(**person,object,from,to**)
prerequisites: **NEXTTO(person,object) & AT (object,from)**
effects: **AT(object,to)**
constraints: **WEIGHT(object) < 300**

In English: for a person to move an object, that person must be next to the object, and if that object is at one location before the moving, it is at another location afterwards. If the person who is to do the moving is not next to the object, the planner may find some plan to make the person next to the object. However, if the object weighs more than 300 pounds, the planner simply gives up.

The importance of plan-based approaches to language has been that as they move toward a more operational view of language, they enable researchers to ask important questions — for example: What does language operate on? How does the object of interaction constrain the phenomenon of language? What is the effect of an utterance? The work of Diane Litman, considered next, began to answer some of these questions.

Case study: Litman. The guiding principle of plan-based approaches is that understanding an utterance consists of relating it to the underlying plans of the speaker. Two limits of early work, however, were 1) that the plans considered were limited to *domain plans*, e.g., catching a train or assembling a machine, and 2) there was no systematic investigation of the particular ways utterances related to such plans. This probably was because the type of conversational interaction focused on was one in which one of the participants was trying to carry out a domain plan, and the other (modeled by the computer) was assisting the other in doing so. In addition, most work concentrated on single utterances, with the result that utterances were mostly taken to be questions or instructions as to what domain action to do next. The obvious limitation is that much of language deals with the interaction itself.

Litman (1985; Litman & Allen, 1987) addressed these problems with one innovation that had several significant consequences. She

introduced *discourse plans*, domain-independent plans that model how utterances manipulate the topic of conversation, for example, by introducing a topic, modifying some previous topic, or correcting some previous topic. Possible topics of conversation are the domain plans of the particular task, as well as previously introduced discourse plans. This scheme requires an explicit model of the ongoing dialogue. The model consists of a stack of all the discourse and domain plans (Planstack) that have been introduced thus far.

The main practical achievements are the ability to handle a variety of different sub-dialogues, to allow direct or indirect interpretations of the same speech act in different contexts, to bring linguistic coherence intuitions and the use of clue words into the plan based approach, and to do all these things with a relatively simple algorithm. The main points of interest for a comparison with interactivism are the move to an explicit representation of the conversational context, the explicit treatment of utterances as operators on this structure, and that speaker and hearer try to keep this structure synchronized, leading to a number of presupposition-like effects (Bickhard, 1980b).

Litman's introduction of an explicit representation of the conversational context and of discourse plans as operators on it constitutes a major step in the evolution of plan based approaches, by allowing the possibility of both participants to manipulate the topic in various ways, including introducing various sorts of sub-dialogues. It also allows a direct comparison with interactivism; indeed, this work can be seen as an instantiating some of interactivism's programmatic statements concerning the role of situation conventions.

A major point of convergence is that both Litman and interactivism view the model of the conversational context (Planstack or situation convention — with its sub-organization of linguistic situation convention: Bickhard, 1980b) as a structure maintained by each participant, which each participant tries to keep synchronized with the models of the other participants. Bickhard (1980b, p. 124) discusses how the need to maintain synchronization leads to a number of presupposition-like effects: "in order to preserve the presumption that some particular interaction is an utterance, in order to preserve the presumption of appropriateness between that interaction and the contextual situation convention, the presumed situation convention may have to be changed." This is precisely what Litman's algorithm does in certain cases. This will always be the case at the beginning of a dialogue: the plans operated on

by the discourse plan will have to be introduced, or, in other words, the utterance, in order to be taken as an utterance, must change the presumed situation convention. Another case would be if the speaker pops the top two plans off the stack, then produces an utterance that continues what is now her top plan. For the hearer to interpret the utterance, he must re-synchronize the Planstack by popping the top two plans.

Another implication of the interactivist argument is that since utterances are operators (on situation conventions), and the result of an operator always depends on the object to which it is applied, the meaning of utterances is inherently context dependent. Litman's work again provides an instantiation of this statement, as each utterance must be interpreted as recognizing a discourse plan which *operates on* (and possibly creates) *the current dialogue context.* The treatment of indirect speech acts provides a good example. An utterance like "Do you know how much a ticket to Austin costs?" is often taken not simply as a query as to the speaker's knowledge, answerable with "yes" or "no," but rather as a request to be informed of the price of the ticket. However, the interpretation is dependent on the context in which the utterance occurs. For example, if we are both travel agents, and I am trying to find out where our knowledge of ticket prices is incomplete, a simple "yes" or "no" will suffice, but if I am a customer and you are a travel agent, I will want you to tell me the price. Litman's algorithm will process the utterance in exactly the same way; however, the Planstack in the two different cases will contain different plans, thus in the first case the direct, and in the second the indirect interpretation will be found.

However, despite the convergences between interactivism and plan-based approaches in general and Litman's work in particular, there are still fundamental differences. First, all AI planning work, including language planning, continues to assume that the object that is acted upon by plan operators is a structure of encodings. Thus, Litman's Planstack encodes a subset of the possible plans that were specified by the programmer as relevant in the system's domain. Therefore, AI plan-based approaches to language illustrate the combination of a transformation approach to language with an encoding approach to knowledge. This combination is logically coherent (Bickhard, 1980b) but severely limited.

Second, the planning approach does not do justice to the dynamic, situated nature of intelligent activity in general and of linguistic communication in particular. Scant attention is paid to the work required

to interpret plans in a particular situation or to the detailed processes through which speakers and hearers strive to achieve mutual intelligibility and to repair communicative trouble. The discussion of Lucy Suchman's work elaborates on these points. They also are touched on in the discussion of Phil Agre's work and, less directly, in the discussion of SOAR.

Awareness of Social, Interactive Nature of Language

Lucy Suchman: From Plans to Situated Action. Lucy Suchman is a social scientist who has developed an ethnomethodological critique of AI planning and human computer interaction systems based on it (1987). She has argued that the notion of planning as used in Artificial Intelligence and Cognitive Science is adequate neither to describe human activity nor to design effective interactive computer systems. We begin by discussing how ethnomethodology approaches the problems of intelligent activity and mutual intelligibility and showing why this approach calls into question basic Artificial Intelligence assumptions. We then present the alternative conception of *situated action.* Finally, we describe the implications of this work on Artificial Intelligence (and related fields) and conclude by relating ethnomethodology to interactivism. Our presentation draws on Suchman (1987) as well as Garfinkel's (1967a) original work and Heritage's (1984) excellent overview.

Ethnomethodology is a relatively new branch of sociology that takes the issue of how people achieve mutual intelligibility (shared understanding) to be the fundamental problem of social science. It argues that shared understanding involves the application of common procedures, rather than access to a common body of knowledge. Garfinkel and other ethnomethodologists have attempted to discover the nature of these procedures both through detailed observation and analysis of natural interaction and through so-called "breaching experiments," which force participants to violate the usual procedures. Suchman's interpretation of one of these experiments challenges some of the fundamental assumptions of Artificial Intelligence in general and the planning approach, in particular.

Garfinkel (1967b) asked his students to report a simple conversation by writing on the left side of a piece of paper what was said and on the right side of the paper what they and their partners understood was being talked about. He progressively imposed more and more

requirements on his students, finally requiring that "I would know what they had actually talked about only from reading literally what they wrote literally ..." (p. 26) At this point, his students gave up, complaining that the task was impossible. What is crucial is *why* the task was impossible: it would be theoretically uninteresting if they gave up only because it would have taken a ridiculous amount of time to write down the vast, but finite body of information that would have let Garfinkel understand what was being talked about only on the basis of what was written. Garfinkel, however, asserts that the problem was that writing down what was being talked about actually extended what was being talked about. That is, the task was not to write down some existing content, but to generate it.

Suchman (1987, p. 33) goes on to argue that background knowledge is not a pre-existing collection of "things that are 'there' in the mind of the speaker but that background assumptions are "generated by the activity of accounting for an action, when the sense of the action is called into question ... [thus] the 'world taken for granted' denotes not a mental state, but something outside of our heads that, precisely because it is non-problematically there, we do not need to think about." This is crucial because Artificial Intelligence approaches, including planning, presuppose precisely what Suchman and Garfinkel argue does not exist — a finite, enumerable list of facts, or, in other words, a knowledge base (see above, on the unboundedness of interactive implicit representation).

The argument thus far is that the explicit representations of background knowledge required by (traditional) AI approaches do not exist. Suchman also focused on plans *per se* and argued that they do not account for the dynamic, interactive nature of intelligent activity. She characterized the AI view of plans as being mental structures that 1) exist prior to activity, and 2) generate activity. She argued that, to the contrary, fluid unreflective activity is the norm, and that plans arise in various ways *from* activity: for example, they are after-the-fact rationalizations or they are used in "breakdown" situations. The role of plans is as *resources to be consulted* rather than *programs to be followed*. Plans are less like models and more like maps.

> plans share with models the function of supporting projections and reconstructions of action. But rather than abstracting action in the strict sense of constructing a homologue of the action's structure, plans are a simplification or sketch of action. Like maps, and like linguistic formulations of action generally, the utility of a

> plan rests on a particular kind of relationship, constructed at the time of its use. Specifically, the usefulness of a plan requires that the actor construct a correspondence between the plan, and his or her actions under some actual circumstances. (Suchman, 1987, p. 46)

If activity is not generated by plans, and mutual intelligibility is not achieved by applying background knowledge to recognize plans, then ethnomethodology has to construct an alternative account. The basic claim is that mutual intelligibility is a contingent, ongoing, social *achievement*, rather than a set of shared assumptions. People never can guarantee common understanding; instead, they assume common understanding as necessary, act in such a way to make these assumptions public, then use social feedback to repair their assumptions where they prove incorrect. The trademark of ethnomethodological research is observation and fine-grained analysis of naturally occurring interactions, directed at discovering the nature of the procedures that people use and the resources they employ to achieve common understanding for-all-practical-purposes.

Studies (see Heritage, 1984, Chapter 8, and Levinson, 1983, Chapter 6, for overviews, Atkinson & Heritage, 1984, for a collection of papers) have revealed that interaction consists of constantly ongoing joint work between the speaker and listener. There is a system of local control for managing the interaction in which resources such as prosody, gesture, gaze, and timing are used by the speaker to gauge whether the hearer seems to be "getting it" and by the hearer to indicate his or her interpretations. One of the most elegant results has been to show that even silence can be a meaningful contribution, i.e., when some substantive comment is expected.

A: I think this paper is going to take the world by storm, don't you?
B: (pause of a few second with no response)
A: or not?

Another important observation has been that troubles of various sorts are ubiquitous in communication and that methods for repairing trouble are a natural part of the local control system.

Suchman applied these theoretical understandings in carrying out an ethnomethodological analysis of people's first interactions with an expert help system that provided instructions in the use of a large, complex copying machine. It had been observed that people reported the

system was too complicated, got confused, and were unable to complete their copying jobs.

Suchman first analyzed the interaction model underlying the design of the expert help system. The "knowledge" of the system was in the form of a set of plans for carrying out the various copying tasks supported by the copier. Users would specify their job by indicating the state of their originals (e.g., bound vs. unbound, 1-sided vs. 2-sided) and the desired properties of the copies. The system would use this specification to retrieve a plan to carry out the specified job. The system then effectively attributed this plan to the user — its job was to instruct the user through the plan step-by-step.

Sometimes this design succeeded, and users were able to complete their jobs successfully. However, Suchman noted that the help system and users had very different relationship to the plans used by the system, and this difference led to serious interaction problems. The plans determined the system's behavior, but the users had to figure out what the plan was from the system's instructions and situational resources. The system and users had very different situational resources. For example, none of the work that the users went through in interpreting referents and action descriptions was available to the system. This asymmetry of resources often caused users and system to have *different definitions of the situation*, and there were insufficient resources to discover and correct these differences, and this in turn led to interaction failures that could not be resolved.

Suchman's critique is relevant to all areas of Artificial Intelligence (and related fields) that deal with interaction. However, natural language research *per se* has paid little attention to her work. The biggest influence has been on the areas of computer supported cooperative work and human computer interaction.

As should be apparent, there are deep convergences between interactivism and ethnomethodology (and the situated action movement, in particular). First, both share a social conception of knowledge, in which knowledge is interactively constructed and maintained. Second, ethnomethodology has a powerful argument that the "background" against which actions are generated and interpreted cannot be a finite set of pre-existing mental structures. It instead points to the world itself as being the background. This parallels interactivist arguments against the inherent limitations of encoding spaces. Third, both interactivism and ethnomethodology see language as a social resource for maintaining

social realities. Fourth, the ethnomethodologist's social achievement of mutual intelligibility **is** the achievement of a situation convention. And fifth, the reflexivities involved both in the processes of such achievement and in the ontology of the social realities achieved (Mehan & Wood, 1975) are specifically modeled by situation conventions (Bickhard, 1980b). Ethnomethodology's detailed empirical analysis nicely complements the theoretical accounts of interactivism.

The major distinction we see between interactivism and ethnomethodology is that ethnomethodology offers *only* social explanations. Although it has strong critiques against AI/Cognitive Science models of psychological processes, it gives no alternative account of underlying psychological processes involved in producing activity. Among other consequences, there is no grounding of any of the presumed representations — no way in which they could have representational content for the system itself. In contrast, interactivism offers an account of psychological processes in terms of goal-directed interaction and the separation of representational function and representational content. Interactivism presents a model of how the social emerges from the psychological (Bickhard, 1980b, 1992a), and it offers a model of representational content.

Winograd and Flores: Applying Heidegger to AI. In his dissertation, Terry Winograd (1971) carried out one of the seminal works in AI language studies, presenting a procedural approach to language understanding. His system, SHRDLU, displayed apparently quite sophisticated understanding. He later did fundamental research in procedural semantics (Winograd, 1976), frame-based representations (Bobrow & Winograd, 1977), and frame-based language understanding programs (Bobrow et al., 1977). However, in 1986, he and Fernando Flores published *Understanding Computers and Cognition*, a Heideggerian inspired critique of Artificial Intelligence and its presuppositions concerning the nature of intelligence, representation, and language. Their work shares with interactivism a focus on the fundamentally social nature of language and a critique of standard AI notions about representation. However, we argue that a residual encodingism in the sources they draw on, Heidegger, hermeneutics, and the work of Maturana and Varela, lead them to the brink of a social solipsism. In addition, they are strongly anti-AI, arguing that the question of whether computers can be intelligent is incoherent. Interactivism, on

the other hand, does not give up the goal of constructing naturalistic models of intelligence.

Winograd and Flores ground their work on the phenomenology of Heidegger. This foundation leads them to reject staples of the "rationalistic" background of cognitive science like "sentences correspond to things in the world" or "knowledge is a storehouse of representations." Our discussion will focus on two aspects of their argument — the "blindness" of AI representations and the social, ontological nature of language — and interpret them from an interactivist perspective.

In Heideggerian analysis, the world is fundamentally experienced as *ready-to-hand* — roughly, experienced *non-reflectively* as a world of possible actions and instruments *ready-to-hand* for those actions. Objects and properties per se emerge only in *breakdown* situations, in which they become *present-at-hand*. In the paradigmatic example, a hammer

> as such does not exist. It is part of the background of *readiness-to-hand* that is taken for granted without explicit recognition or identification as an object. It is part of the hammerer's world, but is not present any more than are the tendons of the hammerer's arm. (Winograd & Flores, 1986, p. 36)

A hammer emerges *as a hammer*, or becomes *present-at-hand* only in a breakdown situation, say when the hammer breaks or slips from the hammerer's grasp or mars the wood. Treating a situation as present-at-hand, i.e., experiencing it terms of objects and properties, creates a blindness, in which one's options are limited by the terms that have been selected (Winograd & Flores, 1986, p. 97).

Applying this analysis to AI programs, what a programmer does is precisely to make a situation present-at-hand, i.e., to define a task domain in terms of objects, properties, and operators. The program's scope is strictly delimited by the representation of the domain that the programmer has supplied it with: it cannot construct new properties or new operators (except, of course, as combinations of those supplied by the programmer). Winograd and Flores point out that the essence of intelligence is to act appropriately when there is no specified definition of a situation in terms of objects, states, operators, and so on. What we recognize as intelligence and creativity has much more to do with the construction of representations of problems than with search in existing representations.

Recent social science work has made the same point (Schön, 1983; Lave, 1988). Going back more than 20 years, a classic AI paper by Saul Amarel (1968) presents a series of representations of the "Missionaries and Cannibals" problem and shows how problem solving is drastically effected by the choice of representation. What is most striking is that all the intellectual work reported in the article was done by Amarel (in inventing and analyzing the various representations), not by his program (in searching within the representational space it was given). It is telling that even now, no AI program can generate its own representations of a problem (unless, of course, a programmer gives it a space of possible problem representations among which to choose; this simply poses the same problem at a higher level).

Winograd and Flores' critique of the blindness of AI representations is a parallel to the interactivist claim that nothing new can emerge from an encoding system (see discussions of encodingism above; also our discussions of SOAR and CYC). However, they do not offer any even remotely technical solution to the problem of representational blindness. For those who (like us) seek models of representation in both humans (and other animals) and artifacts, this is precisely what we want (Bickhard, 1993a).

Winograd and Flores view language as a social tool that serves social functions. It does not transmit information from a speaker to a hearer. There can be no perception-to-cognition-to-communication sequence. Rather, language is a tool that helps people coordinate their actions. It helps to anticipate and cope with recurrent patterns of breakdowns in concernful activity. And, language fundamentally requires commitment, or the ability to engage in dialogue, to expand on and to account for one's actions, in the face of breakdown. This social characterization of language is strongly convergent with interactivism.

Despite the deep similarities between interactivism and the Heideggerian approach espoused by Winograd and Flores, there is at least one crucial difference. As we mentioned above, while Winograd and Flores rightly critique AI representations, they offer no alternative account of how an agent — computational or human — can have epistemic contact in the world. This deficit can be traced back to the later Heidegger, whom it led at least to the very brink of a linguistic solipsism. Winograd and Flores assert that "nothing exists except through language" (p. 68), but aware of their danger, go on immediately to say that "we are not advocating a linguistic solipsism that denies our embedding in a world

outside of our speaking." However, they say that things exist or have properties only by being part of a "domain of articulated objects and qualities that exists in language and through the structure of language, constrained by our potential for action in the world." The first part of this phrase seems a straightforward admission that existence depends on language; the slender branch they cling to comes in the second part — "constrained by our potential for action in the world." Our reading of their work is that they do not have any technical account of what this means. Interactivism has no such problem; it offers an account of epistemic contact with the world as emerging from goal-directed interaction — Bickhard (1992a, 1993a) indicate how this avoids the problems of encodingism, including idealism — as well as how the social emerges from the psychological.

Conclusions

We have traced some of the history and current trends in Artificial Intelligence language studies. We have shown Artificial Intelligence models of language have become increasingly transformational, incorporating views of language as context dependent, knowledge-rich, action- and goal-oriented, and fundamentally social. This has led to great pressure on Artificial Intelligence models of knowledge to evolve to accommodate the kind of processing required to use language. However, Artificial Intelligence models of *knowledge* remain essentially encodingist, thus limiting the scope of possible progress in language studies. In addition, critiques of Artificial Intelligence in general and approaches to language/communication in particular by Suchman and Winograd and Flores have been largely ignored by the language community. This community still sees language as a primarily cognitive, rather than social phenomenon.

12

Learning

Learning is one of the areas in which the programmatic impasse of encodingism encroaches on even the current practical goals of Artificial Intelligence, as well as on the theoretical attempts of Cognitive Science. Encodingism restricts learning by limiting it to the combinatoric space defined by the encoding atoms. Even more fundamentally, it is simply not possible to learn new encoding atoms at all — just as the representational content for new encoding atoms cannot be defined, so also it cannot be learned. Only interactive indications, and their implicit predications, can be learned.

RESTRICTION TO A COMBINATORIC SPACE OF ENCODINGS

One straightforward practical problem is a simple consequence of the fact that any standard symbol manipulation system can at best explore in various ways the combinatoric space of encodings that is determined by that system's base set of encoding representational atoms — encodingism provides at best a representational chemistry, not physics. This is not to deny that, here as well as elsewhere, a great deal of power can be derived from such explorations (Carbonell, 1989), but the limitations are clear and severe. In effect, learning within the encoding approach can be no more than a different principle by which the combinatoric space is explored — e.g., combinatorial variations and selection principles instead of rules of valid or heuristic derivation.

Note that, although it would be quite inefficient, logical derivation rules could be themselves set up as selection principles against which combinatorial variations would be tried — retain a new trial combination only if the new combination is derivable by some rule from already accepted symbol strings — so even standard systems are simply a special case of such a variational and selection process. In this special case, the variations are guaranteed to satisfy the derivational selection principles

because the selection principles are themselves renderable as derivational rules. The limitation for such encodingist variation and selection models is similarly the same as for standard models: No new representations are possible, and, therefore, the encoding atoms must successfully anticipate all possible contingencies in their space of combinations.

Particularly in the case of learning, however, failures of such anticipation are exactly what learning is functionally for — if anticipation succeeds, then learning is not needed. But genuine learning of new encoding representations is foundationally impossible. This is, in fact, the fulcrum for Fodor's argument against the possibility of learning and development (Bickhard, 1991c; Fodor, 1981). These comments require modification with respect to the developments of connectionism and PDP, but, as we shall find, the modifications do not deflect the basic encodingism critique.

LEARNING FORCES INTERACTIVISM

Passive Systems

The central theme of this discussion is that learning is not possible in passive systems. A primary conclusion, in fact, is that the only kind of representational content that *is* learnable is *interactive* representational content. This is certainly consistent with our arguments that interactive representational content is the only kind that can exist, but the approach from issues of learning provides a new argument for that conclusion.

The argument against passivity is, roughly: Learning requires error. Error requires that a system take some action, produce some output, that could be wrong — which thus excludes passivity. Concerning what *can* be learned — only indications of the (internal) consequences of an (inter)action can be in error (for the system itself). Therefore, only indications of internal consequences — interactive representational indicators — can be learned.

The impossibility of learning in passive systems, and the consequent impossibility of learning anything other than interactive representation, is a theoretical result — but one that has immediate practical consequences. In particular, within the fundamentally *passive* epistemic framework of encodingism, naturalistic general learning is impossible. Not *all* systems that make encodingist assumptions, of course, are strictly passive, but the correspondences with the world that are supposed to make elements in the system into encoding representations *are* strictly passive. Encodingist representationality does

not depend on the existence or even the possibility of outputs or interactions. This impacts learning because learning requires error, error for the system, in order for the learning procedure to be invoked, and discovering error is problematic when there are no outputs.

This problematic of error in passive systems has three consequences for learning: 1) Error itself cannot be learned in a passive system. Consequently, passive systems can at best "learn" input functions that satisfy *fixed, built-in* criteria for success and failure on the products of those functions. 2) Error *feedback* cannot be generated without output. Taking an action that does not "work as expected" *is* error feedback. Learning is not possible without such error feedback. Output, by definition, constitutes non-passivity. So, not only can error itself not be learned in passive systems, no learning *at all* can occur in passive systems. Turning then to issues of what *can* be learned, we find: 3) What can *be in error* in a feedback system is only "indicated (internal) consequences of output in these (implicitly defined) circumstances." That is, what can be in error is interactive representational indicators. But, if learning requires error (for the system), and only interactive representational indicators can be in error (for the system), then only interactive representation can be learned. The "encoding" correspondences that might be generated by input processing *cannot* be in error *for a system* — however true it may be that those input correspondences play a *functional* role in detecting or tracking the circumstances for emitting particular outputs, or engaging in particular further interactions.

There are a number of superficially apparent counterexamples to the general claim that passive systems cannot learn. These include: 1) Passive systems with built-in error criteria, such as an internal specification of a loss in chess for a program that plays against itself. 2) Learning with the designer as tutor, as in typical connectionist models. 3) Dedicated error signals, such as pain, and 4) reinforcement learning more generally. Each one of these turns out to be either not passive after all, or of in-principle restricted power in learning. In the course of developing the general argument that "What can be in error is interactive representation, therefore what can be learned is interactive representation," we will visit these seeming exceptions.

Passive systems with Built-in Error Criteria. Passive systems with built-in error criteria can be of practical use in some circumstances in which the nature of relevant errors that a designer wants a system to

"learn" to honor is easier to specify than is the manner of avoiding those errors. So, the designer can specify what counts as error, and the system can learn how to avoid that specification of error. In such a case, a designer might design in a trial-and-error (or heuristic) manipulator of the functions computed on the inputs such that the manipulator — for example, a connectionist back-propagation procedure — would be satisfied only when the designed-and-built-in error criteria were avoided. The designer might, for example, want the inputs to be classified into **X**s versus **Y**s. If the training inputs can be designer-known and designer-pre-classified to be instances of **X**s and of **Y**s, then an error-driven process can in principle be constructed for a "passive" input classifier to "learn" to classify such inputs as **X** instances or **Y** instances. This requires, however, that those *classifications* of **X**s and **Y**s be input to the system along with the **X**s and **Y**s per se, so that the classifications of the system can be compared to the "correct" classifications that have been input. This is in fact the way in which connectionist systems often "learn" (see below). Another example is a chess playing program that plays against itself and learns with respect to designed-in criteria of "win" and "lose." Chess error criteria are clearly easier to specify than are algorithms for chess winning!

In general, however, these models capture at best designer-learning, not general learning. All error criteria must be pre-designed for such a system. This knowledge of what constitutes error is not available to the system, and is not discoverable by the system, except via the ad-hoc-for-the-system designed error classifications. Such knowledge of what constitutes error constitutes a fundamental prescience relative to the general problem of learning in a natural environment. Many of the fundamental problems of natural learning are already "solved" when such error criteria are provided.

More deeply, we will show that even such systems with designed in error criteria are still not really passive — *no* learning is possible in *strictly* passive systems. The "system plus error signal generator" as a joint unit *might* be passive with respect to the external environment (if both are resident in the same computer, for example), but the system itself must emit outputs to the error criterion agent — whether that be a program or some other form — in order to receive error feedback.

Learning with the Designer as Tutor. A slightly different version of this model is the designer-as-tutor. Here, the system computes some function on the inputs, and outputs the result to the designer. The

designer then feeds back, via some built in signals, information about "correctness" or "incorrectness." In such cases, what the system learns is to generate outputs that satisfy the tutor. If the outputs that satisfy the tutor happen to be classifications of the inputs that are in some way meaningful for the tutor, then, in this case too, there might be some practical point in such a system.

But, once again, the error criterion work is being done by the designer/tutor, both in building-in error-feedback signals in the system, and in generating those error-feedback signals. The designer does all the cognitive work of determining what counts as an error. Further, as mentioned, in such a case what the system is actually learning is what outputs to emit under what internal conditions — any correspondence between those internal conditions and categories in the environment are strictly functional or causal, not epistemic. This cannot be a general solution to the problem of learning in natural environments.

Note that even in the case of a designer functioning as tutor, there still has to be some built-in error condition or signal in the system. There must be some way that a feedback of "correct" or "incorrect" can be input to the system, and the system must respond to that feedback with the proper special "learning" activities — e.g., back-propagation in a connectionist net — rather than just processing the signal as just another input like all the other inputs. That special signal, and the special internal response that makes it a special signal, must be built-in to — designed-in to — the system.

In the case of the program playing chess with itself, the outcome of play must be given to the "win-lose" signal generator, and that signal of "win" or "lose" must be fed back to the playing program *as an error signal.* The indication of "win" or "lose" must trigger appropriate learning activities in the program, not just be processed as some further standard input. The "win-lose" evaluator, then, must serve as an internal *tutor*, and the two conditions of "internal error criterion" and "external tutor" turn out to be minor variants of each other.

Similarly, in the case of classifying into **X**s and **Y**s, the correct classifications that are input must be compared to those classifications that are generated by the system, and an internal error signal generated if there is a mis-match. This internal error signal, in turn, must evoke proper learning processes, perhaps back-propagation, and not be processed like an ordinary input.

In all such cases, then, there must be some error condition in the system that triggers learning processes, and there must be some feedback generator that can generate that error condition. Correspondingly, there must be some output to that generator so that the feedback can be evoked. This general outline is required for all learning: learning requires error, and error requires output.

Most fundamentally, learning processes must *operate on* whatever engages in basic system processes — such as input processing, or interaction. Learning *changes* such basic system processes. System processing and learning are related but intrinsically different functions. Learning — adaptive change — must be guided by errors of the processing system, so some sort of evaluation must connect the processing with the learning. The processing must provide output for evaluation, and the evaluative result "error" must evoke learning. There is, however, a more general version of a system error condition than we have considered so far — one that does not require a specific input signal or processing state to be *dedicated* as an error signal or state.

Skepticism, Disjunction, and the Necessity of Error for Learning

The basic issue here is that learning must be responsive to error, so there must be some way to assess error — to generate error signals. We have already discussed several conceptual issues involved in the possibility of error existing for representations and for the epistemic agents that have those representations. Two of these issues are that of skepticism and the more recent disjunction problem. They shift focus from the necessity for error, and the consequent necessity for output, to the related issue of what can be in error, and, therefore, what can be learned.

The skepticism perspective on the problem of error-for-learning interacts interestingly with the disjunction problem. In brief, the skepticism problem is that an encoding system cannot check for error, since any such check is simply a check of one encoding against another instance of the same encoding type. If the first token is wrong — presumably wrong from an observer perspective — then the second will also be wrong. The disjunction problem emerges in considering how to decide whether a factual correspondence between something in the environment and a purported system encoding is correct or not. If a horse on a dark night evokes the "cow" encoding, why doesn't that simply show that the "cow" encoding actually encodes the disjunction "cows or horses

on dark nights"? The fact that the *observer* may be able to classify the "horse on a dark night" instance as an error does the *system* itself no good if the system is supposed to learn to avoid errors. This is in effect an observer perspective on the skepticism-solipsism problem: How can the observer avoid being forced to condemn all epistemic systems to solipsism?

The disjunction problem, then, as a version of the skepticism-solipsism problem, provides a particularization of the learning problem for encodingism. An encoding system, in order to learn, must be able to determine error for the system itself. But, among other kinds of potential error, that implies that the system must be able to solve the disjunction problem *for itself* — an observer or designer "solution" does not address the *system's* problem. But, if all the system is capable of is additional passive encodings, then it can at best recompute its "cow" encoding as a check against the first evocation of that encoding, and that is no check at all — it is circular and will never differentiate between "correct" cow evocations and "incorrect" horse-on-a-dark-night evocations (Bickhard, 1993a).

In a passive encoding system, encodings as input-function correspondences cannot be discovered to be wrong by the system itself. But learning *requires* that such error be distinguishable by and for the system. To be able to transcend the level of designer-provided error knowledge, then, requires that the system be able to determine error for itself, and that is impossible within an encodingism. In this way, the practical problem of learning encounters head-on the philosophical impossibilities and incoherences of encodingism.

Interactive Internal Error Conditions

For an interactive system, however, the *strictly internal* conditions that are indicated as outcomes of indicated interactions constitute internal error criteria. If those internal conditions are not reached, are not induced via interactions with the environment, then an error has occurred. This is an error that is potentially functional for the system itself. In particular, it can invoke learning constructions on the organization of the system. Furthermore, such constructions of new system organization can in principle construct new such internal error criteria. A goal of *avoiding* such an internal error criterion is just a moderately complicated switch (Bickhard, 1993a), and provides no in-principle problems of construction,

given a system that can constructively respond to failures at all (Bickhard, 1973, 1980a; Bickhard & Campbell, in preparation).

It is only with respect to the functional "expectations" that are implicit in the generation of outputs — interactions — that error for a system can be defined, and it is only with respect to such system error criteria that general learning can occur. In the general case, a system can only learn to avoid the errors that it can functionally detect.

Dedicated Error Signals. The claim at this point is that interactive indications of internal outcomes, and only interactive indications of internal outcomes, can serve as error criteria. One possible (and superficially attractive) rejoinder is to postulate something akin to pleasure and pain feedback in an otherwise passive system, with the pleasure and pain contingent on the products of the input processing functions.

This, however, is simply a variant of designer learning. "Pleasure" and "pain" are the designer specialized inputs for serving the error defining function. It is because of the designed specialization of such inputs that the system can differentiate success-failure inputs from any other kind of "just more inputs to be processed." In biology, evolution might serve the function of designer for some degree of this designer-knowledge of what constitutes error. Hence this approach can capture some degree of learning.

There are two points that we wish to make in response to this rejoinder. First, dedicated error signals are a variant of the interactive sense of error-as-failure-of-functional-anticipation. *Strictly* passive input processing systems have no way of classifying some inputs as success and some as error. All inputs are equally simply more inputs to the input processing system. The designer can build-in a *special signal* that constitutes error feedback — that carries "correct" or "incorrect" controlling information to the learning processes — but this now requires *output* from the system in order to evoke the error feedback.

Such a specialized error input signal, however, is just a simple and rigid version of interactive error. A dedicated error *signal* is simply one that the system will always, context independently, respond to by entering an error *condition*, and subsequently switching to whatever the error learning response is designed to be. It is the error condition, and the consequent control flow switching to the learning process, that is crucial. The fact that the system has been designed so that it always enters that error condition upon reception of the "error signal" is just a simple and

rigid version of an internal interactive failure. Dedicated error signals, then, can be useful in some circumstances, but they cannot be a general solution; they are a limited variant of internal interactive error.

A second point is that dedicated error signals require that the designer ensure that the evocation of the error signal does in fact correspond with something that the designer wishes to be taken as error. If the designer-as-tutor is feeding back the error signals, then the designer is providing the error criterion. If the error signal is to be generated in some other way, then the designer must ensure that it gets generated only in appropriate circumstances — in circumstances that *are in error* in whatever sense the designer intends. The designer must build the signal in, and the designer must take care of all of the "semantic" issues involved in what this error signal really "corresponds to." The designer does all the epistemic work. For learning, this is not programmatically better than designer or user semantics for representation.

Reinforcement learning. There is, of course, a natural version of learning from dedicated error signals: reinforcement learning. Pain, for example, can serve as an error signal. Pain can induce a transition into an "error condition" for learning (in addition to reflex withdrawal, and so on). Reinforcement learning, however, at best captures the learning to avoid errors that evolution-as-designer has been able to "learn" about and to provide that foreknowledge to the individual organisms. Such a model of learning intrinsically requires a source of such foreknowledge, and, therefore, cannot be adequate as a general model of, or a general approach to the construction of, learning systems. In addition, the vast majority of human learning is not driven by reinforcement.

Again, the most basic point is that any such designer learning, whether from evolutionary "design" or human design, is still of necessity an interactive learning system. There must be outputs to some generator of error feedback, and an error condition that can be induced by that feedback, and a learning process that is evoked by that error condition. Reinforcement learning, and other forms of designer learning, can be useful and important ways of providing heuristic foreknowledge for further learning *when that foreknowledge is available*, but they are "just" special, less flexible, versions of the general interactive, error-feedback learning framework, and do not introduce any new in-principle considerations.

What Could be in Error?

We have argued that only interactive indications can generate error for the system, and, therefore only such indications can be learned. This has fundamental consequences for models of representation: if representations are to be learned, then they must have some rather close relationship with interactive indications. We argue as follows: What is representational for a system must be capable of having some (fallible) truth value for the system. Therefore, what is representational for a system must be capable of being in error for the system. Therefore, only the implicit predications of interactive indicators can be representational for a system, because only such interactive indicators can be in error for a system.

Learning requires error. Error requires output. But, given such output induced error feedback, *what* is it that is in error — for the system? What evokes an error signal is the emission of *that* output in *those* circumstances; so it is such an "output in such circumstances" that *is* in error. The error in the system, then, is the internal functional connection between the system states and the emission of those outputs with those indicated internal consequences — *not the input correspondences between the system states and the environment*. It is only internal-state to output-processing paths of the system that can be in error and, therefore, it is only the indications of such paths in the functional organization of the system that can *not* be in error. It is only indications of possibilities of interaction that can "succeed" in generating error, and, therefore, that can succeed in *not* generating an error feedback. Consequently, and finally, it is only such indications of potentialities for interaction that can have truth values for the system itself, and, therefore, it is only such indications of potentialities for interaction that can constitute representation — interactive representation.

Error as Failure of Interactive Functional Indications — of Interactive Implicit Predications

In general, any input or input string (or, more generally, flow of interaction) could indicate error *if it were not part of the indicated interaction potentialities.* Error is most fundamentally the entry of the system into internal states that were not indicated as paths of future interaction.

Error is entry of the system into internal conditions that are not among those indicated for the output or interaction engaged in. That is,

error is falsification of the implicit predication about the environment — the predication that is implicit in the indication. An indication of the potentiality of some interaction with its associated possible outcomes *constitutes* an implicit predication to the environment that that environment is of a sort appropriate for the interaction that was indicated. Such predications, in turn, constitute interactive representation. Hence, if those indications are falsified, then so is that predication falsified. Interactive error is error of interactive representation.

Learning Forces Interactivism

Encodingism, by virtue of its intrinsic commitment to *epistemically* passive systems, cannot capture general learning. Learning requires error, and passivity requires foreknowledge of error. Encodingism is committed to *epistemic* passivity because encodings are defined in terms of correspondences with, classifications of, inputs — in terms of the products of functions on inputs. There is nothing in principle to prevent an encoding system from being constructed with outputs, but those outputs will at best be functionally contingent on input derived encodings, and will be totally superfluous to the presumed epistemic content of the presumed encodings. Activity of an encoding system may be desired and built-in by a designer, but it is never required by the nature of encodings per se.

If the possibility of error feedback is built into an encoding system, then the system can (perhaps) learn to satisfy whatever criterion controls the generation of the error signal. But what gets learned in such circumstances is the proper functional control of the emission of outputs — how to emit what outputs in which (internal) conditions so that the error signal is avoided. Input correspondences may well play an important functional role in this, but it is only a functional role, not an epistemic role. Such a system does *not* learn what it's input correspondences are correspondences *with*.

The only aspect of such a system that has a representational character is the internal functional indication that "this" is a condition appropriate for the emission of "that" output, or the engagement in "that" interaction. When that output is in fact emitted, the error generator assesses it and provides feedback, or the interaction proceeds as anticipated or not as anticipated. The system learns to emit proper outputs in proper circumstances; it does not learn input correspondences. It is only such indications of appropriate output or interaction that can

generate error, thus only such indications that can be in error, thus only such indications that can *not* be in error. It is only such indications that can have the fundamental representational property of having a discoverable truth value *for the system itself*. So, it is only such indications that can *be* representations for the system itself. But these are interactive representations, not encoding representations.

Learning cannot be of encodingist correspondences — encodingism cannot model learning. Encodingism, then, focuses on correspondences generated via the processing of inputs to the system — presumed encodings. But the system has no knowledge of the existence of those correspondences, nor of what they are with. If an encoding system attempts to test some "representation" of what such a correspondence is with, it immediately encounters the skepticism problem — all it can do is process inputs once again. A system with outputs and internal error criteria can (in principle) learn to satisfy those error criteria, but what it learns, then, is how to interact in a way that avoids error criteria. Input correspondences play at best a functional role, not a representational role. *Factual* input correspondences, after all, are precisely what are created by functional detection or differentiation processes — there is a correspondence with whatever is detected or differentiated. Encodingism, however, takes those factual correspondences to be *representations*. Encodingism presupposes that the system knows, represents, what it has detected or differentiated; interactivism proposes only that the system knows, or can learn, how to interact further having made such a detection.

Encodingism, then, completely misconstrues what gets learned in a learning system, and, consequently, seriously distorts explorations of how learning occurs, and misleads explorations of how to design learning systems. In this manner, encodingism creates not only a theoretical impasse, but also a *practical* level impasse with respect to the aspirations of Artificial Intelligence and Cognitive Science to model and to build genuine learning systems.

Learning and Interactivism

In sum, we have:

- Only that which can be wrong for the system, can be right for the system.
- Only that which can be right for the system can be representational for the system.

- Representation is, therefore, that which can be wrong for the system.
- Only interactive indication, and its implicit predication, can be wrong for the system.
- Interactive indication and implicit predication *is* interactive representation.
- Therefore, only interactive representation can be representational *for the system*.

Classical encodings, in contrast, cannot be in error for the system itself — what they are to encode cannot be checked by the system: this is the classic problem of skepticism. Correspondingly, classical encodings cannot have representational content for the system itself. If the correspondence cannot be checked by the system, then it cannot be wrong for the system, and therefore cannot be right for the system, and, therefore, cannot be representational for the system. These consequences are just the learning perspective taken on the basic circularity and incoherence of encodingism. Looking backwards down the input flow — to see what is on the other end of the input correspondence — is required for classical input processing encodings to exist, but that "spectatoring" down the input flow requires precisely the representational capabilities that are purportedly being modeled — it is a circular modeling approach, and intrinsically so.

Consequently, it is worth pointing out that, not only is interactive representation the only *learnable* representational content, as well as the only *possible* representational content, but — to shift emphasis on the same point — interactive representation *is learnable*. No encodingism provides any possibility of representational learning at all. Input processing correspondences *can* be designed or defined or trained, but there is no possibility of the system learning for itself what is on the other end of those correspondences — there is no possibility for the system to make those correspondences into genuine epistemic encodings.

COMPUTATIONAL LEARNING THEORY

Explorations of "learning" are often not concerned with issues of representation at all. The "Computational Learning Theory" movement (Angluin, 1992; Valiant, 1984) illustrates this point. This work applies the apparatus of theoretical computer science — analysis of algorithms and complexity theory — to the problem of deriving "learning" algorithms. Rivest and Schapire (1987, 1989), for example, are

concerned with the presence or absence of "learning" algorithms for correctly inferring the state-transition organization of a black box automaton. This is a difficult and interesting problem, but their focus is on existence conditions, not issues of origin or construction, and the required representations are simply designed-in to their actual and hypothetical "learners." In effect, they are asking for successful algorithms and selection conditions for the task of state-transition inference *in presupposed already constructed appropriate encoding systems*. Theirs is a design perspective, and the emergence of representation is not part of their design concerns. Their work illustrates both the power and complexity, on the one hand, of problems and results available in this field, and, on the other hand, the neglectful presuppositions regarding foundational representational concerns. Such neglect at the level of any one, or any proper subset, of particular projects in the field could simply be a manifestation of divisions of labor within the field, but, of course, the encodingist presupposition is inherent in the standard definition of the field, and is not recognized and addressed anywhere in the field.

INDUCTION

A direct focus on problems of learning is to be found in Holland, Holyoak, Nisbett, and Thagard (1986). In their book *Induction*, Holland et al describe a general approach to the problem of induction, explain a number of empirical results in terms of this model, and present several computational models developed within the theory. Several aspects of their presentation are noteworthy.

First, they conceive of induction — construed broadly as the construction of new knowledge — as occurring in service of a system's goal pursuit. Specifically, induction occurs when the system fails in its attempts to achieve a goal. This view is important for two reasons. First, it makes learning a fundamental part of the activity of a system. This contrasts with the common AI conception of learning as distinct from other everyday activities. This leads to models in which language understanding, for example, is a process in which the system does not understand anything that it does not already know. Second, goal-directedness and the possibility of failure are brought in as important aspects of representational activity.

The second important feature is that learning, at least in Holland's classifier system computational model, occurs via a process of variation

and selection. In a classifier system, knowledge is represented as condition-action pairs, consisting of fixed length binary strings. Rules gain strength by participating in successful interactions with the environment and lose strength if they participate in unsuccessful interactions. New rules are created by applying various "genetic operators," such as random mutation and crossover to existing rules. The stronger a rule is, the greater the chance that it will be the parent of a new rule. New rules replace existing weak rules. Thus, successful knowledge (the "strong") is likely to serve as the basis of new knowledge ("reproduce") and unsuccessful knowledge ("the weak") is likely to disappear (to be "selected out").

While the approach of Holland et al has these several attractive qualities, it still is grounded on an unquestioning acceptance of encodingism. Representation is taken to consist of internal elements that correspond to elements in the environment. As with most other models we have discussed, there is no discussion of how factual relationships between external and internal states become known to the system. Although Holland et al employ goal-directedness and goal-failure to drive the creation of new knowledge, there is no appreciation that the very possibility of representational content depends on goal-directedness. At best, such a system learns new condition-action pairings, not new representations.

One symptom of this encodingist assumption is that while they speak of the need for a system to be able to detect new features of its environment, they give no account of how this could be accomplished. Another is that while variation and selection plays a role in classifier systems, it is variation and selection within an already given encoding space. In contrast, interactivist variation and selection occurs with respect to system functional organization, from which representation is a functional emergent.

GENETIC AI

Drescher proposes a computational model of Piagetian sensory-motor development (1986, 1991). According to Piaget (1954, 1962, 1970b), the infant begins with only basic reflexes between sensory and motor systems, and must *construct* knowledge and representation of the world — including, most particularly, knowledge of the permanent existence of objects, even when those objects are not perceivable. This constructive process proceeds through several stages of development,

each building upon and incorporating the preceding, and culminating in the construction of the scheme of the permanent object (Drescher perpetuates a translation error in many of Piaget's earlier works, using "schema" for Piaget's "scheme": "schem***a***" for Piaget has a quite distinct meaning). Piaget's writings are based on fundamentally biological metaphors, not on computational mechanisms, and Drescher proposes an approach to the computational implementation of Piagetian constructivism.

Overview

Novel Representations. One of the most important characteristics of Piagetian development is that it involves the construction of fundamentally new representations. As is by now familiar, this is impossible in standard encoding frameworks. This constructivism has been, in fact, an explicit point of conflict between Piaget and rabid innatist encodingists (Piattelli-Palmarini, 1980).

Drescher does not miss this fundamental point: "A major goal of my research is to elucidate a mechanism for the creation of novel elements of representation." (1986, p. 1) From the interactivist perspective, the emergence of such novel representation is the fundamental aporia of Artificial Intelligence and Cognitive Science, so Drescher's goal, and his success in reaching it, is of central importance.

Drescher's constructive ground is the notion of scheme (we maintain Piagetian usage), and a set of foundational schemes relating innate sensory input conditions and motor output conditions. A scheme is a set of initial sensory conditions — represented by a set of *items* — an action, and a set of resultant sensory conditions — items — *that would obtain if the initial conditions were to hold and the action were to be taken*. The emphasized clause marks the introduction of one of Piaget's most important insights: knowledge is not fundamentally of actuality, but, rather, of potentiality — of possibility (Bickhard, 1988a, 1988b, 1993a) — of *what would* happen *if*. Actuality is "simply" a current location in the network of possible transformations among potentialities (Bickhard, 1980b; Bickhard & Campbell, 1992; see above in the frame problems discussion). Piaget often puts the point in terms of environmental states and environmental transformations: there is essentially no knowledge of a state except in terms of knowledge of how it can transform, and be transformed, into other states. This is the central point in Piaget's notion of the intrinsic "operativity" of knowledge (Piaget, 1977).

Modality. The fundamental *modal* character of knowledge — the fundamental involvement of possibility and necessity — is a direct consequence of the fundamental emergence of knowledge and representation out of interaction — out of interaction *potentialities*. This modality is almost universally overlooked (Bickhard, 1988a, 1988b, 1992a, 1993a; Bickhard & Campbell, 1989).

The focus on relative invariants and stabilities of such interaction potentialities — physical objects in particular — obscures this underlying modality of representation in favor of an apparent representational focus on, and limitation to, *actuality* — actual objects. However *ontologically* it may be true (or not) that objects are generally stable in their actuality, *epistemologically* the ability to function in terms of and to represent such stability is a complex constructive achievement out of a fundamentally modal base of interactive potentialities.[19] From an epistemological perspective, an object ***is*** a relatively stable organization of the various interactions that could — *potentially* — be performed with that object.

Furthermore, in children, modalities are initially *undifferentiated* — exactly as would be expected from the underlying modal interaction character of knowledge and representation — and only slowly do children become able to differentiate actuality from possibility from necessity (Bickhard, 1988b; Piaget, 1987). This is in direct contradiction to the standard approaches to knowledge representation and reasoning strictly in terms of extensional (encoded) actualities, to which modal considerations must be *added*, if taken into consideration at all.

Drescher, in endorsing Piaget's notion of scheme, has captured this most overlooked aspect of knowledge and representation — its intrinsic modality. He has also established a deep grounding convergence with interactivism.

On this ground of basic sensory-motor schemes, Drescher posits several processes by which new schemes can be constructed. These include differentiating schemes into separate new ones in terms of differing initial conditions and differing resultant conditions, with several

[19] There has in recent years been a flurry of work in cognitive development that claims that children are born with innate concepts and theories of such things as objects. At best, such claims would establish that the essential constructions mentioned above have occurred in evolution, rather than occurring in infancy. The claims, in other words, do not have any effect on the basic logical point concerning constructivism. On the other hand, we find serious conceptual and methodological flaws in much of this work. This is not the opportunity to pursue those criticisms, so we will simply register this demurral. To repeat, however, the basic logical point concerning constructivism is not influenced one way or the other by such claims of innatism.

computations of statistical properties of past experience with actions serving to initiate and control such new constructions.

Synthetic Items. The most intriguing process of construction occurs when a regularity of action sometimes occurs and sometimes not, with no conditions currently available to the system that differentiate when that regularity will occur and when it will not. In such circumstances, a "synthetic item" is constructed to represent the *unknown* presumed differentiating conditions. As further learning takes place regarding what schemes yield "manifestation" of such a synthetic item and what sorts of schemes are contingent on it, it progressively becomes *known* in the sense of the system being able to act in accordance with the properties of whatever it is in the environment that that item does in fact "represent."

Drescher shows how such synthetic items could be expected to yield the construction of object permanence — how a synthetic item could come to represent an object even though it were hidden from view. In such a case, the synthetic item "represents" the conditions under which actions that depend on the hidden object will succeed, and differentiates from the conditions under which those actions will not succeed. That is, the synthetic item "represents" the object, even though hidden.

Convergences

There is a most remarkable convergence with interactivism here — due, of course, largely to the mutual overlap with Piaget (Bickhard & Campbell, 1989). This extends even to small aspects: Drescher's "items" are close to the "indicators" of Bickhard (1980b), and his notion that "Designating an event as an action in its own right is a way of abstracting above the details of how to achieve that event." (1986, p. 9) is similar to the notion of "cofunctionality" of Bickhard (1980b). The notion of "scheme" in interactivism is open to much more complexity than Drescher's "single action" definition, but the underlying point is similar, and *complex organizations* of Drescher's schemes could capture more of the notion of "scheme" in Bickhard (1980b).

Differences

There is, nevertheless, a basic difference between genetic AI and interactivism. In critical respects, this carries over a difference between interactivism and Piaget (Bickhard & Campbell, 1989). One perspective on this difference is to note that Drescher's items *always* "represent"

environmental conditions. Drescher is very careful with the fact that that representation relationship is for him alone, and not necessarily for the system — the system must *construct* whatever schematic power those items will have for it — but, nevertheless, each item is presumed to be in correlation, in correspondence, with some environmental condition or conditions, even if totally empty for the system itself. This contrasts with interactive indicators which serve *solely* to indicate the outcomes of some interactions — external *or* internal — that might be used to further indicate the potentialities of other interactions. There is no apriori reason why such indicators will necessarily have *any* coherent correspondence with the environment, rather than, say, with relevant strictly-internal machine conditions. (It can be argued, in fact, that *language* involves the emergence of the external relevance of initially strictly internally relevant such indicators, Bickhard, 1980b.) And if there should be any such environmental correspondences, the system does not thereby represent anything about them. Any representational function that does occur does so in terms of the indications of further potential interactions.

One important place where this difference makes a difference is in terms of Drescher's "synthetic items." These represent unknown environmental conditions that differentiate when a scheme will function as predicted and when it won't. The system then constructs further schemes around such synthetic items, and, thereby, tends to fill out their place in the overall organization of potential schematic actions. First, this requires a distinct construction process for synthetic items; second, its success requires that there in fact *be* such environmental differentiating conditions. In the interactive model, the function of differentiating further system activity is the fundamental function, and reason for construction, of *all* indicators. If that further differentiation also turns out to have potential representational significance, then that having-of-significance may be discovered in further constructions, but there is no necessary *posit* of such representational significance required for the initial construction of any such differentiator of system activity. Interactivism allows for the *discovery* of representational significance, should it exist, and the constructive *emergence* of its representational content, rather than requiring that representational significance apriori, and, therefore, restricting constructions to possible representational contents. *Indicators* have a fundamentally functional nature, out of which the function of representation can emerge; *items* have a primarily

representational nature, around which and toward which schematic function is constructed.

Encoding Correspondences. There is a subtle, vestigial encodingism here — and in Piaget (Bickhard, 1988a; Bickhard & Campbell, 1989). Representation is *constructed* in Drescher's model, but it is constructed around correspondences with environmental conditions. Furthermore, for Piaget, it is constructed as action *structures* that are correspondent — isomorphic — with the *potential* transformations that are in the world. That is, although making the vital move from actuality as the locus of representation to a ground in the modalities of action, Piaget nevertheless ends up with a *structuralist*-correspondent encoding notion of representation at that much deeper level of the realization of the modality-of-action-character of knowledge and representation (Bickhard, 1988a; Bickhard & Campbell, 1989; Campbell & Bickhard, 1986). With his always environmentally correspondent items, and his schemes of single, explicit actions, Drescher is also committed to this structuralist notion of representation in the larger sense — although his discussion does not explore this level.

From the interactivist perspective, Drescher's model does not have the notions of interactive differentiation nor of the emergence of representational content in the further indications of interactive possibilities. His model does not have the critical notion of implicit definition — though his discussion of the relationship between intension and extension begins to touch on it. He avoids many of the problems of encodingism by his care and attention to the fact that the items "represent" at least initially, only to him. He does not, however, provide a full account of how they could come to have representational content for the system — he does not provide an account of representational emergence.

Drescher (1991) also argues forcefully for the importance for learning of indications of the *outcomes* of actions and interactions.[20] He rightly notes that general learning is impossible without such indications: it is only in terms of errors in such outcome indications that learning can be guided. Furthermore, Drescher models the construction of synthetic

[20] Though, again, his indications of outcomes are of sensory-input outcomes — correspondences between the system and the sensory environment — not internal state outcomes per se. This is not only a vestigial encodingism on the representational level of analysis — as it is for Piaget — it also potentially problematic on a strictly functional level of analysis; see the "Kuipers' critters" discussion for a brief exploration of the difference.

items on the basis of such outcome indication successes and failures. There is, therefore, a notion of pragmatic error, and a construction of representational items on the basis of such pragmatic error. This is a close convergence with the interactive model (Bickhard, 1980b, 1993a).

The representational content of those items, however, is still given in terms of correspondences to environmental conditions, not in terms of such pragmatic successes and failures. Pragmatic error and representational error are still distinct in kind, however closely related in the processes of construction. Drescher has recognized the importance of pragmatic error for learning, but has not recognized the emergence of representational error, thus representational content, out of pragmatic error. In the interactive model, in contrast, representational error is *constituted* as a special kind of pragmatic error. Representational content is emergent in the interactive indications among the implicitly defined conditions: failure of the pragmatic indications *is* falsity of the implicit predications. The distinction here between Drescher's model and the interactive model is essentially that mentioned in the discussion of pragmatics between 1) the function of representation being anticipation, and 2) representation being the function of anticipation.

Drescher belongs to a small company of AI and Cognitive Science researchers who recognize the fundamental ground of representation in action, and incorporate the intrinsic modal character of that ground. Furthermore, his principles and processes of the construction of new schemes out of old ones on the basis of experience — which we have only touched upon — are interesting and of fundamental importance to any adequate constructivist model. The degree of convergence, and the depth of convergence, between genetic AI and interactivism is intriguing.

Constructivism

Drescher's recognition of the importance of constructivism is, in the current context, one of the most important characteristics of his model. Only with a genuine constructivism that allows for the emergence of novel representations can Artificial Intelligence and Cognitive Science begin to understand and capture learning and development — and this holds for strictly practical AI as much as for programmatic Artificial Intelligence and Cognitive Science. Drescher's recognition of the basic importance of constructivism, and his contributions toward an understanding of constructive processes, are a welcome change from the simple shuffling of encodings.

Some Intrinsic Connections. It should be noted that there are strong logical connections among several points about representation and development. The recognition of the ground of representation in *action forces* a *modal* ground of representation — the *organization* of *potential* action or interaction is the essential focus. In turn, the ground in the modal action organization forces a constructivism: environmental *potentialities* cannot impress themselves on a passive system — those potentialities don't *exist* — and even environmental *actualities* cannot passively induce a functional system organization in a system. This stands in contrast to the seductive notion that actualities — objects or events or patterns of events — can impress themselves, inscribe themselves, as correspondence encodings on a passive, tabula rasa mind. This tight logical progression of action, modality, and constructivism is inherent in genetic AI, Piaget's genetic epistemology, and interactivism alike (Bickhard & Campbell, 1989). It will similarly impose itself on any other approach that recognizes the basic emergence of representation out of action.

13

Connectionism

Connectionism (Waltz & Feldman, 1988b) and Parallel Distributed Processing (Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986) are variants of an approach to cognitive phenomena that has, with good reason, stirred much excitement and controversy. For our purposes, the differences between the two variants are not material. The general distributed approach has a number of distinct differences from standard compositional or symbol manipulational approaches to cognitive phenomena, some of which constitute definite advances and strengths, and some of which may be relative weaknesses (Clark, 1993; Norman, 1986). We will explore both. The central point to be made, however, is that connectionism and PDP approaches are just as committed to, and limited by, encodingism as are compositional, or symbol manipulational, approaches (though in somewhat differing ways). The "representations," the symbols, of a connectionist system are just as empty *for the system* as are those of any standard data structure (Christiansen & Chater, 1992; Van Gulick, 1982); a connectionist approach does not escape the necessity of a user semantics.

OVERVIEW

A connectionist system is, first of all, a network of nodes. Various pairs of nodes are connected to each other by a directional and weighted path, and some nodes are connected directly to sources of input from the environment. For a given system, the topology of the connections among the nodes and from the environment is fixed. *Activations* are received along the connections from the environment, which activate the directly connected nodes, which in turn activate further nodes via the weighted paths among the nodes. A given node receives activations from all upstream nodes in accordance with the activation levels of those nodes and the weights of the paths of transmission involved, and acquires a

degree of activation resulting from those inputs in accordance with various built-in rules. It then sends its own activation down further connections in accordance with their weights. Often the nodes are organized in layers, with each layer sending activation levels along connections to the next layer. Most commonly, there are three layers of nodes: an input layer, a "hidden" layer, and an output layer whose ultimate activation levels constitute the intended output of the net of nodes — a three layer feed-forward net. Loops of connection-and-node paths are possible, however. Activation thus flows through the network, possibly with feedback loops, until, in some circumstances, a stable pattern of activation of the nodes is achieved.

A fixed connectionist system is, therefore, specified by three sorts of information. The first is the graph of the nodes and directed connections among them and from the environment. The second is the set of weights on those connections. The third is the rules by which the nodes determine their resultant activations given their input activations. Both the graph and the weights, together with the relevant rules for setting node activations from inputs, determine a space of possible patterns of activation of the nodes, and a dynamics of the possible changes in activation patterns, forming paths or trajectories through the space of possible activation patterns. In general, this space will be a space of vectors of possible node activation levels, and the dynamics will determine trajectories of possible movement through paths of activation vectors.

In important cases, those dynamics will determine a set of regions or points in that space of possible activation patterns in which the activation patterns will remain stable — the trajectories of activation patterns that exist in or enter such a stable region do not exit. Furthermore, there will be a tendency for initial activation patterns, upon receipt of environmental inputs, to "settle" into those determinate stable points or regions. In other words, the areas of stability will each have their own "catch basin," "drainage basin," or "region of attraction" such that any patterns of activation in one of those basins will dynamically move through the space of possible such patterns *into* the corresponding stability. Ideally, the regions of attraction for the points of stability will correspond to desired categories of inputs; they will differentiate the space of possible inputs into those categories that yield one particular pattern of resultant stable activation versus another.

The directed graph of a particular connectionist network is, in general, fixed, as are the various rules of functioning of the system. That graph and those rules together with the set of weights determine the space and dynamics of the activation patterns of the system. They determine the trajectories of the activation patterns within the overall space of possible such patterns, and they determine, as an aspect of that dynamics, which regions of activation patterns, if any, are stable, and what the regions of attraction to those stabilities, the "catch basins," will be. They determine which input patterns will stabilize, if at all, at which points of stability.

As described thus far, a connectionist system is a *differentiator* of input activation patterns in terms of the resultant stable activation patterns. The excitement concerning connectionism and PDP derives from the fact that the weights of the system are themselves not fixed, and that it proves possible in some cases to adjust the weights according to well defined rules and error correction experiences so that particular desired differentiations of input patterns are obtained. Such adjustment of the weights is then interpreted as learning — learning of the desired differentiations, categorizations, of the input patterns.

The space of the (vectors of) *possible weights* of a connectionist system is a second-order space of the dynamics of the system. It is the space in which the dynamics of the learning processes occur. Each point in this *weight space* determines its own corresponding dynamics (and stabilities, and differentiations) of the *activation space* — that is, each *point* in the weight space determines an *entire dynamics* of the activation space. In effect, each point in the second order dynamic space of weights determines, and has associated with it, an *entire* first order dynamic space of activations.[21] The dynamics of this *weight space*, the second order space, in turn, are determined by the directed graph of the system together with the "learning" rules and experiences. Movement of the system in this second order "learning" space of weights, thus, constitutes movement in a space of possible activation dynamic spaces, each with its own particular dynamic attractors and attracting regions, and, therefore, with its own particular differentiating properties. Much attention is given to designing system graphs, learning rules, and tutoring experiences that can yield activation dynamic spaces with interesting "categorizations" of inputs. It is this possibility of learning, of the dynamic acquisition of

[21] A fiber bundle in differential geometry — see Chapter 15.

designer specified categorizations of inputs, that has sparked so much excitement within and about the field.

STRENGTHS

The most important advantage of PDP "symbols" is that they are "emergent." They emerge as stabilities of attraction regions in the spaces of activation patterns and dynamics, and "correct" such stabilities and dynamics are generated by appropriate points in the space of possible connection weights. Such appropriate points in the weight space are sought by the "learning" rules in response to the instructional experiences. This possibility of emergence is explicitly *not* present in the typical "symbols by syntactic definition only" approach, in which the best that can be attained are new combinations of already present "symbols."

A second advantage that is claimed for the PDP approach is that the computations of the new activations of the nodes are logically parallel — each new node value can, in principle, be computed at the same time as each of the other node values. This provides the power of parallel computation in that potentially large networks could engage in massive computation in relatively short elapsed time. Massive parallelism and relatively short computational sequences are taken to capture similar properties and powers of processes in the brain (Churchland, 1989).

A related advantage of the connectionist approach is the sense in which the differentiating stable patterns of activation are intrinsically distributed over the entire set of (output) nodes. This distributivity of "representation" originates in the same aspect of connectionist net architecture as does the parallelism of computation mentioned above — each of the multiple activation nodes can in principle also be a node of parallel computation — but the distributivity yields its own distinct advantages. These are argued to include: 1) as with the parallelism, the distributed nature of the "representations" is at least reminiscent of the apparently distributed nature of the brain, and 2) such distributivity is held to provide "graceful degradation" in which damage to the network, or to its connection weights, yields only a gradual degradation of the differentiating abilities of the network, instead of the catastrophic failure that would be expected from the typical computer program with damaged code. This is reminiscent of, and argued to be for reasons similar to, the gradual degradation of hologram images when the holograms are physically damaged.

The phase spaces of the PDP approach introduce a general approach to doing science that, even when abstracted away from the particulars of PDP, is extremely important, and all too rare in studies of mental phenomena. The spaces are spaces of the potentialities of the overall conditions of the network — the space of possible activation vectors, in this case — and, thereby, contain the possible dynamics of the system. Understanding the system, then, is constituted as understanding the relevant phase space and the dynamically possible trajectories of the system's functioning within that space. This is the almost universal approach in physics, but explicit consideration of spaces (or other organizations) of *potentialities* within which dynamics can be modeled is rare in psychology, AI, and related disciplines (Bickhard & D. Campbell, in preparation; van Gelder & Port, in press).

This phase space approach highlights several of the powerful characteristics of PDP models. The space of activation patterns of a PDP network is a space of the *intrinsic dynamics* of the system, not a space of (encoded) information that the system in some way makes use of. It is like an automaton in which the states form a smooth surface (differentiable manifold), the state transitions are continuous on that manifold, and the state transitions intrinsically move "downward" into local attraction basins in the overall manifold. The space, then, does not have to be *searched* in any of the usual senses — the system dynamics *intrinsically* move toward associated differentiating regions of stability.

Viewing connectionist systems in terms of modeling their (weight space adjustable) intrinsic dynamics, instead of in terms of the classical programmed informational manipulations and usages, is an additional perspective on both their distributed and their parallel nature. Because the activation space is the space of the possibilities and possible dynamics of the *entire system*, and because nothing restricts those dynamics to any simply isolable subspaces (such as would be equivalent to changing just one symbol in a data structure), then any properties of that space, such as input-differentiating dynamically-attracting regions of stability, will necessarily be properties distributed over the whole system and dynamically parallel with respect to all "parts" of the system.

Such an overall system perspective could be taken on standard symbol manipulation systems, but it would not be necessary to do so in order to understand the relevant properties, and, in fact, it would lose the supposed information separated out into the "pieces," the "particles," of the system that are taken to be "symbols" — the "symbols" would all be

absorbed into the overall state of the system, and the distinction between program and representation would be lost. (Furthermore, the representations *could not be recovered.* Their *internal functional* properties could be recovered, say, in an equivalent register machine, but the "aboutness," the representational properties, would have to be re-defined by the user or designer). The most interesting properties of a PDP system, on the other hand, *cannot* be analyzed at any more particulate level.

The connectionist approach in terms of the phase space of the entire system also gives rise to several other possibilities. In particular, the sense in which the system dynamics intrinsically *arrive at* the relevant activation stabilities, the sense in which the system dynamics *constitute* the movement into the differentiating stabilities — instead of the system dynamics being distinct from those differentiators and, therefore, the dynamics having to engage in a search for or construction of them — provides for the possibility of several different interpretations of that dynamic "movement toward dynamic stability."

If the input pattern is construed as a *sub*pattern of its corresponding overall stable pattern, then that system can be interpreted as being engaged in *pattern completion.* If the input pattern is interpreted as being a *component* pattern of the attracting stability, then the system can be interpreted as engaging in *content addressing* of the stable pattern. If the input pattern is interpreted as being one of a pair of *joint* patterns instantiated in the stable pattern, then the system can be interpreted as engaging in *pattern association.* Note that if the input pattern and the stable pattern, whether interpreted as subpattern-full pattern, component pattern-subsuming pattern, or paired pattern-joint pattern, were separate *components* of the system — e.g., separate "symbols" or "symbol structures" — then any such completions, content addressing, or associating will have to be accomplished via standard searches, manipulations, and look ups.

The power of the PDP approach here is precisely that the input patterns and the stable patterns are both simply points in the space of the system dynamics, and that the system dynamics *intrinsically* move into the attracting regions — the system does not have to explore, reject, manipulate, draw inferences from, combine, or in any other way consider any alternative paths or dynamics, it simply "flows downhill," "relaxes," into the low point of the catch basin that it is already in. That PDP

systems can apparently do so much with such a natural and simple principle of dynamics is still another of its appeals.

A further power of PDP differentiators relative to symbol manipulation encodings is that, in standard cases, since the dynamics of the system are bound to ultimately enter some attractor or another, the system will ultimately differentiate all possible input patterns. These differentiations have an intrinsically open aspect to them in that, although the system may be trained to "correctly" differentiate certain paradigmatic training input patterns, its response to novel patterns is not necessarily easily predicted from its classifications of trained patterns. Its differentiations will depend on particulars of its organization that have neither been specifically designed nor specifically trained. This yields an intrinsic power of *generalization* to the differentiations of a PDP network. Such generalizations are *not* intrinsic to standard symbols, and are difficult to program. Specific such generalizations may, of course, count *against* specific PDP systems on the ground that the observed system generalizations are not those of the learner which is intended to be modeled, e.g., Pinker's critique of a PDP past tense "learner" (Pinker, 1988). The general power of the *potentiality* for such generalizations, however, is a distinct advance in this respect over predefined encodings. The openness of such differentiations provides, in fact, some of the power of a selection function or of indexical representation, and, therefore, some of the power of variables and quantifiers with respect to issues of generalization (Agre, 1988; Bickhard & Campbell, 1992; Slater, 1988).

WEAKNESSES

The PDP approach, however, is not without its own weaknesses. Most fundamentally, the primary advantages of PDP systems are simultaneously the source of their primary weaknesses. On one hand, the emergent nature of connectionist differentiations transcends the combinatoric restrictions of standard symbol manipulation approaches. Any model that is *restricted* to combinations of *any* sort of atom, whether they be presumed representational atoms or any other kind, intrinsically *cannot* model the *emergence* of those atoms: combinations have to make use of already available atoms (Bickhard, 1991b). The classical approach, then, cannot capture the emergence of input pattern differentiators, while PDP approaches can.

On the other hand, while the combinatoricism of the standard approach is a fatal limitation for many purposes requiring fundamentally

new, emergent, representations — e.g., learning, creativity, etc. — it has its own strengths, some of which the PDP approach arguably cannot capture (Pinker & Mehler, 1988; Fodor & Pylyshyn, 1988). In particular, the combinatoric atomism of the standard symbol manipulation approach allows precisely what its usual name implies: the manipulation of separable "symbols." This creates the possibility of (combinatoric) generativity, componentiality, high context and condition specificity of further system actions and constructions, a differentiation of representational functions from general system activity, a "lifting" of representational issues out of the basic flow of system activity into specialized subsystems, and so on. All of these can be of vital importance, if not necessary, to the modeling of various sorts of cognitive activity, and all of them are beyond the capabilities of connectionist approaches as understood today, or at least can be approximated only with inefficient and inflexible kludges. A restriction to combinatorics dooms a model to be unable to address the problem of the emergence of representations, but an inability to do combinatorics dooms a system to minimal representational processing.

There is, however, a conceivable "third way" for connectionism: the processing of connectionist "representations" that is in some way sensitive to "representational" constituents without there being any syntactic tokens for those constituents. This would be akin to the processing of logical inferences directly on the Gödel numbers of predicate calculus sentences — without unpacking them into predicate calculus strings — or some more general methodology of distributed representation (van Gelder, in press-b). Whether such a strategy is sufficiently powerful, however, remains to be seen — but initial indications are quite interesting (Bechtel, 1993; Clark, 1993; Niklasson & van Gelder, 1994; Pollack, 1990; van Gelder, 1990; van Gelder & Port, 1994; see also Goschke & Koppelberg, 1991).[22] The space of possible "other ways" for connectionism to handle such problems has not been exhaustively explored (Clark, 1989, 1993; Bechtel, 1993).

In general, the freedom of combination of symbolic atoms that provides such apparent computational advantages to classical approaches does not logically require that the relevant "representations" be constructed out of such atoms: it only requires that the space of possible

[22] There are also interesting questions about the supposed naturalness with which *classical* symbol systems can capture the systematicity that human beings in fact manifest (van Gelder & Niklasson, 1994).

constructions have in some manner the relevantly *independent* dimensions of possible construction, and of possible influence on further processing. Invoking constructive atoms is one way to obtain this independence of constructive dimensions — separate dimensions for each atom, or each possible location for an atom — but the critical factor is the independence, not how it is implemented (Bickhard, 1992c). Again, however, it is not yet clear whether connectionist systems per se can make good on this.

Absent any such "third way," however, connectionist systems do have a distinct disadvantage with respect to the systemic constructions and manipulations of their "representations." Put simply, symbol manipulation approaches have no way to get new "representations" (atoms), while connectionist approaches have no way of doing much with the "representations" that they can create. Of course, in neither case is there any possibility of *real* representations, for the systems themselves.

There are other weaknesses that are of less importance to our purposes, even though they could constitute severe restrictions in their own right. Perhaps the most important one is that the learning rules and corresponding necessary tutoring experiences that have been explored so far tend to be highly artificial and inefficient (e.g., back-propagation, Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986; Rich & Knight, 1991). It is not clear that they will suffice for many practical problems, and it *is* clear that they are *not* similar to the way the brain functions. It is also clear that they cannot be applied in a naturalistic learning environment — one without a deliberate tutor (or built-in equivalent; see the discussion of learning in passive systems above). (There are "learning" algorithms that do not involve tutors — instead the system simply "relaxes" into one of several possible appropriate weight vectors — but these require that all the necessary organization and specification of the appropriate possible weight vectors be built in from the beginning; this is even further from a general learning procedure.)

PDP research thus far has focused largely on idealized, simplified problems. Even in such a simplified domain, however, the learning processes have at times demonstrated impressive inefficiencies, requiring very large numbers of training experiences. This limitation of experience with PDP networks to simplified problems, plus the clear inefficiencies even at this level, combined with the general experience of standard symbol manipulation approaches that scaling up to more realistic problems is usually impossible, at least at a practical level, has yielded the

criticism that PDP and connectionism too will find that, even when models work for toy problems, they do not scale to real problems (Rumelhart, 1989; Horgan & Tienson, 1988). This is a criticism *in potentio*, or in expectation, since the relevant experience with PDP systems is not yet at hand.

A limitation that will not usually be relevant for practical considerations, but is deeply relevant for ultimate programmatic aspirations, is that the network topology of a PDP system is fixed. From a practical design perspective, this is simply what would be expected. From a scientific perspective, however, concerning the purported evolutionary, the embryological, and, most mysteriously, the developmental and learning origins of such differentiators, this fixedness of the network topology is at best a severe incompleteness. There is no homunculus that could serve as a network designer in any of these constructive domains (see, however, Quartz, 1993).

ENCODINGISM

The deepest problem with PDP and connectionist approaches is simply that, in spite of their deep and powerful differences from symbol manipulation approaches, neither one can ever create genuine representations. They are both intrinsically committed to encodingism.

The encodingist commitments of standard approaches have been analyzed extensively in earlier chapters. The encodingist commitments of PDP approaches follow readily from their characterization above: PDP systems can generate novel systems of input pattern differentiators, but to take these differentiat*ing* activation patterns as *representations* of the differentiat*ed* input patterns *is* to take them as encodings.

Note that these encodings do not look like the "symbols" of the standard "symbol" manipulation approach, but they are encodings nevertheless in the fundamental sense that they are taken to be represent*ations* by virtue of their "known" correspondences with what is taken to be represent*ed*. Standard "symbols" are encodings, but not all encodings are standard "symbols" (Touretzky & Pomerleau, 1994; Vera & Simon, 1994). Both model representation as atemporal correspondences — however much they might change over time and be manipulated over time — with what is represented: the presumed representationality of the correspondences is not dependent on temporal properties or temporal extension (Shanon, 1993).

The PDP system has no epistemic relationship whatsoever with the categories of input patterns that its stable conditions can be seen to differentiate — seen by the user or designer, not by the system. PDP systems do not typically interact with their differentiated environments (Cliff, 1991), and they perforce have no goals with respect to those environments. Their environmental differentiations, therefore, cannot serve any further selection functions within the system, and there would be no criteria of correctness or incorrectness even if there were some such further selections.

A major and somewhat ironic consequence of the fact that PDP systems are not interactive and do not have goals is that these deficiencies make it impossible for connectionist networks to make good on the promise of constituting real learning systems — systems that learn from the environment, not just from an omniscient teacher with artificial access to an ad hoc weight manipulation procedure. The basic point is that, without output and goals, there is no way for the system to functionally — internally — recognize error, and, without error, there is no way to appropriately invoke any learning procedure. In a purely passive network, any inputs that might, from an observer perspective, be considered to be error-signals will be just more inputs in the general flow of inputs to the network — will just be more of the pattern(s) to be "recognized." Even for back-propagation to work, there must be output to the teacher or tutor — and for competitive "learning," which can occur without outputs, all the relevant information is predesigned into the competitive relationships within the network.

In other words, connectionist networks are caught in exactly the same skepticism-solipsism impossibility of learning that confounds any other encodingist system. No strictly passive system can generate internally functional error, and, therefore, no strictly passive system can learn. Furthermore, even an interactive system with goals, that therefore might be able to learn something, will not be legitimately understood to have genuine "first person" representations so long as representation is construed in epistemically passive terms — as merely the product of input processing — such that the interactions become based on the supposed already generated input encodings rather than the interactions being epistemically essential to the *constitution* of the representations. It is no accident that all "learning" that has been adduced requires designer-provided foreknowledge of what constitutes error with regard to the processing of the inputs, and generally *also* requires designer variation

and selection constructions and designer-determined errors (or else already available designer foreknowledge of relevant design criteria) within the space of possible network *topological* designs to find one that "works."

This point connects with the interactive *identification* of representational content with indicated potential interactions and their internal outcomes — connectionism simply provides a particular instance of the general issues regarding learning that were discussed earlier. It is only with respect to such strictly internal functional "expectations," such contents, that error for the system can be defined, and, therefore, only with respect to such contents that learning can occur, *and, therefore*, only out of such functional "expectations" that representation can emerge. Representation must be emergent out of some sort of functional relationships that are capable of being found in error by the system itself, and the only candidate for that is output-to-input potentialities, or, more generally, interactive potentialities. Representation must be constructable, whether by evolution or development or learning; construction requires error-for-the-system; and the possibility of error-for-the-system requires indications of interactive potentialities. In short, representation must be constructed out of, and emergent out of, indications of interactive potentialities.

In a passive network, however, or any passive system, any classification is just as good as any other. There is no, and can be no, error for a system with no outputs. It is only via the interactions of the *deus ex machina* of the back-propagation (or other "learning" system) with the omniscient teacher that connectionist nets have given any appearance of being capable of learning in the first place.

PDP systems are, in effect, models of the emergence of logical *transducers*: transducers of input categories into activation patterns. But the complexity and the emergent character of the "transduction" relationship in a PDP network does not alter the basic fact that the system itself does not know what has been "transduced," nor even that anything like transduction, or categorization, has occurred. All relevant representational information is in the user or designer, not in the system.

A significant step forward in this regard is found in Jordan & Rumelhart (1992). In particular, they generalize the connectionist architecture to an interactive architecture, with genuine input and output *inter*action with an environment, and a number of interesting capabilities for learning about that environment — learning how to control

interactions in that environment — and their model contains goals. They have also built in the necessary topology for the system to be able to generalize from previous experience to new situations. However, the logical necessity of such interactivity is not addressed; the goal-directedness is an adjunct to the focus on learning to control interactions and interaction paths, and is therefore also not elaborated with respect to its logical necessity; the topology involved is generated by the use of continuous variables in the fixed architecture of the system, and does not allow, for example, the construction of new spaces with new topologies (Bickhard & Campbell, in preparation); and there is no consideration of the problem of representation. Given these unrecognized and unexplored potentials, we conclude, similarly to our position with respect to Brooks and his robots, that Jordan and Rumelhart have accomplished more than they realize.

The emergent character, the phase space dynamics, the distributivity and parallelism, are all genuine advances of the PDP approach, but they do not solve the basic problem of representation. They do not avoid the incoherence problem: basic "encoding" atoms in an encodingist approach, even if they are emergent in a connectionist system, still have no representational content for the system itself, and there is no way to provide such content within the bounds of the modeling assumptions. The emergent regularities of connection between differentiating activation patterns and the input categories that are differentiated are *dynamic regularities* of the non-representational, non-differentiator *functioning* of the overall system; they are not *epistemic* regularities nor *epistemic* connections. If **factual** regularities between inputs and resultant system conditions were all that were required for representation, then every instance of every physical law (not to mention chemical, biological, ecological, meteorological, physiological, and social laws, and so on — and even accidental such factual regularities) would constitute an instance of representation, and representation would be ubiquitous throughout the universe. More than correspondence is required for representation, and correspondence is not only not sufficient for representation, it is not even necessary (e.g., "unicorn" or "Sherlock Holmes" or a hallucination). From the perspective of interactivism and its critique of encodingism, for all their differences and comparative strengths and weaknesses, PDP and connectionist approaches are still equally committed to an unviable encodingist perspective on the nature of representation.

CRITIQUING CONNECTIONISM AND AI LANGUAGE APPROACHES

Like connectionism, approaches to language are dominated by encodingist presuppositions. Our discussion of language models did not emphasize this encodingism as much as we have with regard to connectionism; a question arises concerning why. Approaches to modeling language have proceeded on the basis of encodingist presuppositions, and thereby, in our view, fail from the beginning. The development of the field, however, has manifested more and more insights concerning language that are convergent with the interactive view — in spite of those encodingist presuppositions. Our discussion of language models, therefore, focused most on the major developments in this field, which do tend to convergence with interactivism, though we also point out the problematic assumptions involved.

Connectionism, on the other hand, does not just presuppose an encodingism, it makes fundamental claims to have solved representational problems that standard computer models can not solve. The claim to fame of connectionism rests on, among other things, its claims regarding the ability to learn new representations. The fact that those representations are still encodings, and are therefore subject to the encodingism critiques, is far more central to the way in which connectionism presents itself and is known than is the case for typical language models. It seems appropriate, then, to emphasize more this encodingism critique when assessing connectionism.

On the other hand, connectionism offers deeper similarities to the interactive approach — but these similarities are in terms of the dynamic possibilities of recursive connectionist nets (see below), not in terms of the representational claims made by connectionism. Such dynamic possibilities are being explored, but no one in the connectionist field is at this point making a "connection" between such dynamic possibilities and the nature of representation. The closest to be found are some approaches proposing net models that learn dynamic interactions, but there is no representational claim at all in these, and some other approaches that want to make use of the net dynamics to capture dynamic processing of representations, but the alleged representations supposedly being processed in these are standard encodingist "representations."

We find the connectionist movement, then, to be a convergent ally in many respects — particularly with respect to its explorations of parallelism, distributivity, and phase space dynamics — though seriously misguided in its conceptions of representation. We explore below some

of the most intriguing convergences in connectionist research. These convergences emphasize dynamics, and modify or eschew standard connectionist representational assumptions.

IV

SOME NOVEL ARCHITECTURES

14

Interactivism and Connectionism

INTERACTIVISM AS AN INTEGRATING PERSPECTIVE

For all their relative weaknesses and their common commitment to encodingism, the strengths of symbol manipulation and PDP approaches are not to be denied. If interactivism were unable to capture their respective modeling strengths, that would constitute a serious deficiency of the interactive approach. In fact, however, interactivism constitutes a perspective within which the strengths of both "symbol" manipulation and connectionist approaches can be integrated and their weaknesses transcended. We will first explore some senses in which interactivism can suggest some integrating architectures, and then turn to some architectures motivated strictly by the interactive model.

First, as adumbrated in the above discussion, what is in fact emergent in a PDP system is not a full representation, but rather an implicit differentiator. It is a system whose final internal states serve to differentiate input conditions — implicitly — and it is emergently so in that there are no component nor ancestral differentiators out of which the particular differentiating system was constructed or from which it grew. This is just a differentiator in almost exactly the sense of an interactive model.

The caveat "almost exactly" derives from the fact that a PDP system is — so far as its representations are concerned — strictly passive, while an interactive system, in general, cannot be. There are no outputs from the PDP system to its environment which affect subsequent inputs during its dynamic run into an attracting region. Many interesting things can be differentiated with such passive systems, but not everything. An *active* interactive system can only have greater power in such a task of implicit differentiation, and some tasks of differentiation, for any finite system, will require such interactive power. For human beings, one

example would be the various physical manipulations involved in qualitative analysis in chemistry — the end result is still a detection or differentiation, but, since we do not have direct passive transducers for, say, iron, we must perform something like a brown-ring test in order to detect iron. Categories constituted intrinsically as temporally extended — e.g., an animal or other object in motion — and categories with intrinsic temporally extended access — e.g., checking out the rooms in a new house, tracking the animal in motion — intrinsically require something more than atemporal passive reception.

Interactive systems must have more than just implicit differentiators in order to have full representations. Implicit differentiators per se are representationally empty — they are precisely "empty symbols." They acquire representational content in the usages that the system makes of its differentiations in its further interactions. These usages constitute further connections to the environmental conditions that have been implicitly differentiated — "this implicitly differentiated environmental category is appropriate to that further interactive activity." There is still no true representation, however: there is no sense in which these connections to further activity can be right or wrong.

This is the point at which the goal-directedness — the anticipation of subsequent internal states — of the system is essential to true representations. The further system interactions indicated by the connections to the implicit differentiations will tend to be appropriate to and competent for the accomplishment of the goal — or not. That is, those indications of interactive characteristics of the differentiated environments will tend to be correct or not. It is this general success or failure of the indication from an implicit-differentiator to a further-means-to-the-goal that provides a truth value to the basic connection. It will in general be true or false (or a partial truth) that these implicitly defined environments will respond well toward that goal in interaction with these indicated further interactive systems (Bickhard, 1992c, 1993a, in preparation-c).

Not only are PDP systems not interactive in their differentiations, then, they cannot have any representational content for their differentiated conditions since they are not interactive and, correspondingly, not goal-directed. Correspondingly, they cannot model error for the system, and, again correspondingly, they cannot model natural learning from an environment.

On the other hand, interactivism not only provides the ground for the construction of derivative encodings, that can then provide the powers of differentiation and specialization of representational computations, it contains arguments that there will be strong selection pressures for the construction of such derivative encodings (Bickhard & Richie, 1983). Within an interactive framework, then, both the basic power of representational emergence, and the power of representational manipulation, are available. Derivative encodings, however, are not identical to encodings as viewed from within the standard encodingist perspective (Bickhard & Richie, 1983). They do have, nevertheless, the appreciable advantage of being possible.

Hybrid Insufficiency

It should be noted that an interactivist integration that includes both PDP type emergent differentiators and secondary, derivative encodings that can be manipulated, and thus includes the respective powers of each of the approaches, is not just a hybrid of a PDP system with a symbol manipulation system. In particular, a prima facie obvious hybrid would be to take the differentiations provided by a PDP network, or some set of such networks, as constituting and providing the basic encoding elements or element types for a symbol manipulation system built "on top." Such hybrid forms can be very interesting and powerful for some purposes (Honavar & Uhr, 1994), but they still do not have any intrinsic interactions — certainly not *epistemically* intrinsic — and similarly no epistemically intrinsic goals. They simply accept the user-understood encodings from the networks and provide them *as* user-understood encodings to the symbol manipulations. Such a hybrid could combine some of the respective powers of each approach, e.g., the openness of differentiation of PDP networks with the manipulation capabilities of standard programming models, but it does *nothing* to overcome their common error of encodingism. Symbol manipulation approaches and connectionist approaches share that error of encodingism (Honavar, in press); combining the approaches does not address or overcome it. The increased capabilities that would be provided by such a hybrid might be important for some practical design tasks, but such a hybrid does not overcome the fundamental programmatic impasse of representation.

A still further extension of such a hybrid would be for the outcomes of symbol manipulations to control actions of the system in its

world: input, then, would be differentiated into categories — by a connectionist net, say — which would evoke symbols, which would be processed, which would control action. The addition of action, clearly, increases the potential capabilities and usefulness of such a hybrid even further. But, so long as the representationality of such a system is taken by scientists or engineers to be constituted in its input to symbol relationships, rather than in its output-to-further-input relationships, the analysis or design process would still be caught in the encodingist impasse. Such an encodingist understanding of such a system would make it difficult to extend the system design further to include, for example, a general form of learning. The flow of input to system to control of action is critical to representation, but the crucial relationship is circular — a model of representationality must close the circle by including the flow from action to subsequent input — and merely sequential flows do not suffice. Representation is fundamentally a matter of anticipation, and *not* a matter of a system being a retrospective spectator back down its input sequence — somehow "seeing" what is on the other end of the causal input sequence.

SOME INTERACTIVIST EXTENSIONS OF ARCHITECTURE

We have pointed out that an interactivist approach can integrate the strengths of emergent differentiation from PDP approaches and the freedoms of manipulation and combination of symbol manipulation approaches *and* it solves the problem of emergent representation — of representation for the system itself. Simply, the emergent differentiations afforded by connectionist networks are precisely a special case of one of the functional aspects of interactive representation, and interactive representations, once emergent, can ground secondary, derivative encodings that can be available for the sorts of symbol manipulations involved in programming approaches (should such manipulations be needed, which will not always be the case given the other powers of the approach).

There are yet additional properties and possibilities of the interactive approach that further differentiate it from standard connectionist and programming approaches, and from any simple hybrid.

Distributivity

PDP differentiations, in their most interesting variety, are intrinsically distributed over the relevant nodes of the system. This

distributivity is a system level distributivity *underneath* the emergent differentiating activation patterns. That is, nothing about the *differentiating* properties of those attractor activation patterns requires that they be distributed patterns; PDP properties would be just the same if some single final non-distributed atoms were associated with each differentiating pattern, and the ON_OFF status of those atoms were taken to be the "encodings" — frequently, in fact, this is exactly the design used. Another perspective on this point is that the distributivity of PDP models is an implementational distributivity. Nonetheless, such distributivity — in its various distinguishable forms — has considerable power and important manifestations (Haugeland, 1991; van Gelder, 1991, in press-b).

In contrast, there is a distributivity in interactive representations that is epistemically intrinsic — that could not be altered without destroying the fundamental character of the interactive representations. This is a functional distributivity, rather than just an implementational distributivity, involved in the webs of indications of further potential interactions (in the situation image) given an environmental differentiation. Representation is emergent precisely in those functional indications of further potentialities, and those indications are intrinsically distributed within and across the ***organization*** of such indications. That is, interactive representation is intrinsically *relational*, and, therefore, necessarily distributed over organizations of such relations.

The representational content of an interactive representation is constituted in the organization of such webs of indications, and these webs cannot be rendered non-distributed without reducing them to some collection of particulate indications — without eliminating the intrinsic relations. Rendering such a web of relationships as particulate atoms — as representational encodings — necessarily removes the relationships that constitute the web. Such a reduction would be possible in a design sense, but it would make impossible any further constructions of additional such indications — further distributed parts of the web, additional representational content. Additional constructions become impossible because the *relational context* within which new indicative relations could be constructed has been destroyed in moving to the independent particulate indications, and interactive representation is intrinsically constituted as context embedded relational indications. With no relational context, there can be no new indications constructed within such a context.

It might be thought that such relationships could simply be built into the encoding version as encodings of the relationships. But that assumption encounters, among other problems of unboundedness, the frame problems. The web of relationships is implicitly defined by the apperceptive processes, and it is in principle unbounded.

Because the web of indications would not be present in an encoding reduction, that web could not be added to. That is, the representational contents that are constituted by such webs could not be elaborated or expanded. This impossibility of the expansion of indicative "content," in turn, would define the encoding representations in terms of that *fixed* content that they were originally given in the reduction of the original web, rather than in terms of the expandable indicative power of the environmental differentiation — and a representation defined in terms of its content is an encoding. In such an encoding system, in a by now familiar point, new structures of encoding atoms could be constructed, but new encoding atoms could not be. Once the functional indicators are dirempted from the functionally relational action system into particulate encodings, the indications lose their interrelationships via that action system — they become context independent encodings — and new representations cannot be emergent in new relational organizations of the action system. Interactive representations are not just *contingently* functionally distributed, they are *intrinsically* functionally distributed — they are intrinsically *relational.*

A Prima Facie Incompatibility. There is a prima facie incompatibility between interactivism and connectionism with regard to learning: one seems to be a discrete process and the other continuous. Learning in an interactive model must in its logical non-prescient limit involve variation and selection (D. Campbell, 1974). The presence or absence of "selection out," however, would seem to be an all or nothing process, while the relevant "learning" in PDP systems involves adjustments of weights, generating a continuous manifold. The incompatibility disappears, however, once it is recognized that, if variation and selection processes are viewed at a constructive unit level, they are all or nothing, but if they are relevant at a level of *populations* of such units — e.g., modulation relationships among populations of neurons — then all or nothing variations and selections *within* the population determine precisely weights or proportions or ratios at the *level* of the population. More deeply, variation and selection processes will honor whatever structure there is in the space within which the

variations and selections occur. If the variations and selections are with respect to a continuous structure — e.g., trials (whether heuristic or random) and "selections out" of weight vectors forming a continuous manifold — then that continuous structure will be intrinsically honored.

Metanets

Another sense in which an interactive architecture offers extensions of connectionist architecture has to do with the rigidity of the connectedness topology in the typical connectionist net. Recall that interactive architecture is an architecture of oscillatory systems modulating each other's activity (see the Turing machines discussion above, and Foundations of an Interactivist Architecture immediately below). Modulation strengths between various parts of an oscillatory system could be constructed (or set or modulated) by higher level interactive goal-directed systems, thus constructing and even interacting with — manipulating — interactive system network organization. Such organization, then, would not be fixed. An approximation to this could be constructed by taking the *nodes* of one PDP network to be the *connections* of a lower level network, and the activations of the nodes of the higher level network to be the weights of the lower level system. In this way, one network could affect the connectivity of a different network (Lapedes & Farber, 1986; Cowan & Sharp, 1988; Pollack, 1991, see also Quartz, 1993). A more flexible version of this would construe the weight of a given connection to be the product of the "normal" weight that is determined by the "learning" of the first level system times a zero or one depending on a threshold criterion concerning the activation of the relevant node of the higher level system: in this manner, the higher level network would set *only* the connectivity of the lower level network (0 or 1), and the learning would proceed within that connectivity "normally." In any case, these variants all at best capture a passive version of the more general interactive higher order constructions of interactive systems. Conversely, there is no problem in principle for interactive subsystems to affect the connectivity of other subsystems — to modulate the topology of modulation relationships.

15

Foundations of an Interactivist Architecture

We shift now to the interactive view, in which the basic system dynamics are constituted as modulations among oscillatory processes. Recall that oscillatory processes are necessary in order to capture the timing aspects of interactivism, and modulatory relationships among oscillatory processes provide a form of functional relationship that is of greater than Turing machine power.

From the interactive view of architecture, both intrinsic dynamic *topologies* and intrinsic interactive *timing* are critical. Connectionist networks, unlike Turing machines, do manifest an intrinsic topology in their dynamic spaces, but, similar to Turing machines, they do *not* have intrinsic timing: the phase space is not an oscillator space.

In the interactive view, connectionist weights might be roughly reconstrued as modulatory strengths and connectionist activations reconstrued as oscillation properties, perhaps oscillatory frequencies (in, for example, neural ensembles, Thatcher & John, 1977). However, although connectionist properties can be reconstrued within the interactive perspective, the reverse does not hold. Aspects of the interactive view, such as the nature of the modulations or the effects of differences and alterations in the oscillatory media (e.g., the neural micro-architecture), are not so readily construed from within the connectionist framework.

MacKay (1987) presents an interesting hybrid. His "node structure" theory has some features of a hybrid between connectionism and interactivism: it involves both node activations and node oscillations, and it emphasizes an interactively flavored integration of perception and action. But MacKay accepts standard notions of representation, and has no particular function in his model for modulation. MacKay's nodes can oscillate, and his model is among the few that recognizes the

ubiquitousness of oscillatory timing, but there is no sense of general control relationships being constituted as modulatory relationships among oscillations. In fact, with the shift to an interactive oscillatory *and* modulatory architecture, a number of interesting new properties and possibilities emerge.

THE CENTRAL NERVOUS SYSTEM

Oscillations and Modulations

An architecture supporting oscillations and modulations makes very strong connection with a number of aspects of brain functioning that both connectionist models and standard models ignore (Bickhard, 1991e). Consider first the oscillatory nature of neural functioning. Endogenous oscillations — non-zero baselines — are intrinsic to many neural processes at both the individual cell and network levels. That is, there is intrinsic ongoing oscillatory activity, not just in response to any other "outside" activity (Gallistel, 1980; Kalat, 1984; Kandel & Schwartz, 1985; Thatcher & John, 1977). This is incompatible with the strict reactivity of the switches or commands in classical symbol manipulation AI and with the strict reactivity of the nodes in connectionism. There is no architecturally local endogenous activity in these perspectives, only reactions to inputs. Furthermore, the potential temporal complexity of oscillatory processes stands in stark contrast to the simplicity and reactive temporal fixedness of switches or "levels of activation." Such endogenous oscillatory activity might seem to make sense for motor control (Gallistel, 1980), but should be irrelevant to cognition according to contemporary approaches. Yet oscillatory activity is ubiquitous in perception and in action (MacKay, 1987), and endogenous oscillatory activity is ubiquitous throughout the central nervous system.

Similarly, modulations among such oscillatory processes are ubiquitous, and there are multifarious such modulatory relationships. Oscillations of neurons or networks modulate those of other cells or networks (Dowling, 1992; Thatcher & John, 1977). Various messenger chemicals — peptides, among others — modulate the effects of other transmitters. In fact, such modulation is ubiquitous (Bloom & Lazerson, 1988; Cooper, Bloom, & Roth, 1986; Dowling, 1992; Hall, 1992; Siegelbaum & Tsien, 1985; Fuxe & Agnati, 1987). Furthermore, a significant population of neurons rarely or never "fire" (Bullock, 1981; Roberts & Bush, 1981). Instead, they propagate slow-wave graded ionic potentials. Such variations in potential will affect ionic concentrations in

nearby extra-cellular spaces, or modulate the *graded* release of neurotransmitters, and, thus, effect the modulations of (modulatory) activity that are occurring via synaptic junctions — a modulation of modulations (Bullock, 1981; Fuxe & Agnati, 1991b).

In fact, there is a broad class of non-synaptic modulatory influences (Adey, 1966; Fuster, 1989; Fuxe & Agnati, 1991a; Nedergaard, 1994; Vizi, 1984). A central example is that of volume transmitters, which affect the activity of local volumes (local groups) of neurons, not just the single neuron on the other side of a synapse (Agnati, Fuxe, Pich, Zoli, Zini, Benfenati, Härfstrand, & Goldstein, 1987; Fuxe & Agnati, 1991a; Hansson, 1991; Herkenham, 1991; Matteoli, Reetz, & De Camilli, 1991; Vizi, 1984). For example, the administration of L-dopa, a precursor of the transmitter dopamine, for Parkinson's disease makes little sense within the strict synaptic cleft model. There is a loss of neurons that normally produce dopamine, and L-dopa increases the dopamine production of those neurons that remain. If this increased production of dopamine affected only the neurons to which the remaining reduced number of dopamine producing neurons were synapsed, then it would simply hyperactivate — hypermodulate — the *reduced* neural network. Instead, the dopamine seems to act as a volume transmitter and increases and modulates the activity of whole local populations of neurons, and, therefore, their networks (Vizi, 1984; Changeux, 1991; Fuxe & Agnati, 1991b; Herkenham, 1991). Similarly, dopamine producing neural grafts can have positive effects even without specific innervations via general regulatory functions of elevated dopamine rather than patterned input (Changeux, 1991; Dunnett, Björklund, Stenevi, 1985; Herkenham, 1991).

Chemical Processing and Communication

One conceptual framework for understanding such influences, a framework that is broader than the classical threshold switching model of the neuron, derives from recognizing that synaptic neurotransmitters are strongly related to hormones (Acher, 1985; Emson, 1985; Scharrer, 1987; Vizi, 1984, 1991). Hormones may be viewed as general information transmitting molecules — modulatory molecules — that exhibit an evolution from intracellular messengers to neurohumors to neurohormones to endocrine hormones (Hille, 1987; Turner & Bagnara, 1976). A number of molecules, in fact, serve *all* such levels of function in varying parts of the body. In an important sense, paradigm

neurotransmitters are "just" extremely local hormones — local to the synaptic cleft — that affect ion balance processes (among others). But not all of them are so localized. There appears to be a continuum ranging from the extreme synaptic localization of some transmitters to the whole body circulation and efficacy of some hormones, with local hormones and "volume" neurotransmitters in between.

Consider first the "slow" end of this continuum. Here we note that the larger volume effects are clearly also longer time scale modulations. In the brain, such slower and larger volume modulations can constitute modulations of the already ongoing modulations among neural oscillations and within neural networks, even to the point of reconfiguring the effective neural circuitry (Dowling, 1992; Hall, 1992; Iverson & Goodman, 1986; Koch & Poggio, 1987). Consider now the "fast" end of the spectrum. At this extreme of space and time considerations we find gap junctions which transmit electrical changes from cell to cell virtually instantaneously, with *no* mediating transmitters (Dowling, 1992; Hall, 1992; Nauta & Feirtag, 1986). Slow wave potential oscillations can also function via direct ionic influences, without mediating neurotransmitters, but with larger volume and longer time scale effects. In addition, these may control the graded release of transmitters rather than all-or-none release patterns (Bullock, 1981; Fuxe & Agnati, 1991a).

Modulatory "Computations"

It should be recognized, in fact, that insofar as the classical nerve impulse train serves primarily the function of transmitting the *results* of more local graded interactions (graded and volume transmitter and ionic interactions) over long distances (Shepard, 1981) — that is, insofar as neural impulse trains, oscillations, are the *carriers* of the *results* of the local graded and volume processes — it is these *non*-classical processes that prove crucial. We might also expect impulse oscillations to *participate* in such volume processes (Bullock, 1981). Neither the "switches" nor the "gates" of standard paradigms can make any sense of such processes. This perspective of multiple forms of modulation, multiple spatial characteristics and temporal characteristics of modulations, differential affinities, varying layers of modulation and meta-modulation, and so on, quickly becomes extremely complex, but it remains extremely natural in principle from the oscillatory-modulations perspective of the interactive architecture.

Oscillatory and modulatory phenomena form a major framework for the architectural organization of the central nervous system. In particular, the multiple sorts of modulatory relationships, with their wide variations in volume and temporal modulation effects, permeate the entire brain. Some systems propagate influences primarily via axonal impulses, e.g., the visual tract (Koch & Poggio, 1987), while others influence via slow wave potential movements. Local and volume transmitters seem to coexist throughout many, if not most, areas of the brain (Fuxe & Agnati, 1991a; Vizi, 1984). Some synapses have been found to release multiple chemicals that serve multiple levels of modulatory function from the same synapse. The key differentiations, again, are those of 1) time and volume and 2) meta-modulations on underlying modulations. Superimposed on them are the differentiations of modulatory influences even within the same local volume via differentiations in synaptic and volume transmitter molecules. Differential sensitivities to differing transmitters yields differential sensitivities to differing modulating influences even for neighboring, even for intertwined, neural networks (Koch & Poggio, 1987).

The Irrelevance of Standard Architectures

None of this makes any sense — it is all utterly superfluous — from standard contemporary perspectives. Neither classical symbol manipulation nor connectionist approaches provide any guidance in understanding such phenomena. This irrelevance of standard approaches has led to the recognition that a major conceptual shift is needed in order to be able to even begin to understand the functioning of the brain (Bullock, 1981; Freeman & Skarda, 1990; Pellionisz, 1991). A shift is required not merely from sequential to parallel processing, but from local processing to volume, or *geometric*, processing. These geometric architectural variations in kinds of modulatory influences available within the overall system — with their range in both spatial and temporal characteristics and in transmitting chemicals — are *exactly* what should be expected from within the interactive oscillatory-modulatory architectural perspective.

Standard approaches, both connectionist and symbol manipulation, simply have *nothing* to say about such phenomena. They can at best be interpreted away, in the familiar manner, as implementational issues — beneath the important levels of functional analysis, which are supposedly captured by symbol manipulations or

connectionist nets. We have argued that such architectures are not, and cannot be, mere issues of implementation. Such oscillatory and modulation phenomena are — and logically must be — the basic form of architecture for any viable, intelligent, or intentional system.

A Summary of the Argument

In summary of the architecture arguments to this point, we have:

- Contemporary conceptions of and approaches to representation are not only wrong, they are at root logically incoherent. They universally assume that representation is some form of correspondence (isomorphism) between a representing element or structure and the represented, but they do not and cannot account for how any such correspondence is supposed to provide representational content, "aboutness," for the animal or agent itself. They are universally analyses from the perspective of some *observer* of the animal or agent, and, therefore, are intrinsically incapable of accounting for the cognitive processes and capacities of such an observer per se.
- Representation is emergent from *action*, in roughly Piagetian or Peircean senses, rather than out of the processing of inputs, as in the dominant forms of contemporary Cognitive Science. Representation is constituted as organizations of indications of potentialities for interaction. Epistemic agents must be active and interactive; it is not possible for passive systems to be epistemic systems.
- Representation is constituted in several forms of implicit definition. The easy unboundedness of implicit definition relative to standard explicit and discrete conceptions of representational elements and structures, as in the frame problems, is another fundamental inadequacy of standard approaches and corresponding superiority of the interactive model.
- Action and interaction requires timing. This is not just "speed" since timing can be either too late *or* too early, but, rather coordinative timing of actions and interactions. This is in contrast, again, to standard Cognitive Science in which certain forms of correspondences are supposed to constitute representation, and such correspondences are logically

atemporal in their representational nature, no matter that they are created, destroyed, and processed in time.

- Timing cannot be modeled within Turing machine theory, nor, therefore, within any model or architecture that is equivalent to Turing machine theory. Turing machine theory captures *temporal sequences* of operations in formal processes, but nothing about the steps in those sequences constrains their timing. The first step could take a year, the second a microsecond, the third a century, and so on without affecting the mathematical or formal properties of Turing machines. We are interested in our physical instantiations of Turing machines, computers, being as fast as possible, of course, but this is a practical concern of elapsed time, not a concern of timing.
- Timing requires clocks, and clocks are essentially oscillators.
- A single clock driving everything is not a viable architecture from an evolutionary perspective, since any changes in overall architecture would have to involve simultaneous changes in the clock connections in order to be viable. Such simultaneity, and myriads of instances of such simultaneity throughout the evolution of the nervous system, is, for all practical purposes, of measure zero.
- An alternative architecture is to use clocks — oscillators — as the basic functional unit, and to render functional relationships among them in terms of modulations among oscillatory processes.
- Any change in basic functional architecture in this framework is intrinsically also a change in the timing architecture — the two are identical.
- This is at least as powerful as standard Turing machine architectures: a limiting case of the modulation relationship is to modulate On or Off — that is, a switch is a trivial limiting case of modulation, and a switch is sufficient for the construction of a Turing machine.
- It is in fact more powerful than Turing machines in that timing is now intrinsic to the functional nature of the architecture and its processes.
- Modulation is more general than discrete functional relationships, such as switches, and can at best be asymptotically

approximated by unbounded numbers of such discrete relationships.

- Oscillations and modulations as the basic forms of functioning are much closer to the actual processes in the central nervous system, and account for many properties that must remain at best matters of mere implementation from either standard symbol manipulation or connectionist perspectives.
- Examples of oscillatory and modulatory properties include:
    1) The basic endogenous oscillatory properties of both single neurons and of neural circuits;
    2) The large populations of neurons that never "fire," but that do modulate the activity of neighboring neural circuits;
    3) The wide temporal and spatial range of variations in forms of modulations in the nervous system, such as gap junctions at the fast end and volume transmitters, that diffuse through intercellular fluid rather than across a synapse, at the slow end;
    4) The deep functional, embryological, and evolutionary connections between the nervous system and the hormonal system.
- The nervous system is, in fact, exquisitely designed for multifarious forms of modulatory and meta-modulatory relationships among its oscillatory processes, and this is precisely what is shown to be necessary by the "representation emerges from interaction, which requires timing" argument.

Whereas both information processing and connectionist perspectives are used for modeling, and both make claims for the nature of, the architecture and functioning of the nervous system, neither of them can account for ubiquitous basic functional properties of the nervous system. The interactive model does account for those properties — predicts them as being necessary, in fact — and, thus, is a much richer and more powerful framework within which to explore and model the nervous system than any other framework currently available.

Interactivism, then, offers a candidate for the "major conceptual shift" that is needed for understanding of the functioning of the brain (e.g., Bullock, 1981; Freeman & Skarda, 1990; Pellionisz, 1991). Noting the ubiquitous and necessary oscillatory and modulatory functional relationships in the brain, however, does not constitute a *model* of *brain* functioning — we have not mentioned anything about brain anatomy, for

example — but it does pose a set of architectural *constraints* on any adequate model of brain functioning, and a set of constraints that is not assimilable within symbol manipulation or connectionist approaches. The interactive model is the only model available that makes sense of what we know of the oscillatory and modulatory manner in which the nervous system actually works and, thus, provides guidance for further explorations.

PROPERTIES AND POTENTIALITIES

Oscillatory Dynamic Spaces

We turn now to a few illustrative explorations of some of the properties and potentialities of such architectures. First, note that the oscillations in any particular part of the system can be rendered as a vector of Fourier coefficients of that oscillation. For a temporally complex oscillatory process, this vector of Fourier coefficients will itself undergo change through some resultant trajectory in the Fourier space. Modulations between such oscillatory processes will constitute control relationships on the temporal developments, the trajectories, of the Fourier vectors in the controlled oscillatory processes. Modulations constitute control relationships between the Fourier trajectory of one oscillatory process and that of another.

A particular part of the overall system could be oscillating in multiple frequencies simultaneously. If the modulatory relationships were differentially sensitive to differing parts of Fourier space, as modulations tend to be, then a single physical part of the system could in principle be active in many functionally distinct parts of the Fourier space — functionally relevant subspaces — simultaneously. The differing regions of oscillation space would sort themselves out in their modulation effects with respect to what sorts of Fourier space differential sensitivities existed in the modulat***ed*** parts of the system.

If such a functional organization were approximated by an automaton, then each functional region of the Fourier space would constitute a distinct state (or transition diagram), and modulation relationships would constitute inducements of, or constraints on, or, in general, control influences on the state transitions (Fourier space trajectories) in the modulated subsystems. The possible states in such a system would have a natural topology in terms of their locations within the Fourier space — they would not be the nominal discrete states of classical automata theory. Such an organization would yield a system in

which each physical part could simultaneously instantiate multiple functional states — the different states in differing regions of the Fourier space. Nonlinearities within oscillatory parts of the system, and in the modulation relationships between them, would interconnect these Fourier-separated continuous automata.

The functional states in such a system would be organized as equivalence classes with respect to which physical part of the system they were instantiated within, and the modulatory *connection* relationships would be consistent among such equivalence classes — that is, oscillatory processes within a given *physical* system part would connect with, and, thus, potentially modulate, the oscillatory processes within the *physically connected* physical system parts. For example, multiple oscillatory processes in single brain regions could simultaneously modulate — control — multiple other processes so long as they were all in neurally connected brain regions.

This view raises the functionally relevant dynamic space to a level even more abstracted from the physical instantiation than does the PDP approach. Oscillations can occur in many parts of a Fourier space simultaneously in a single physical oscillator, or oscillatory medium — unlike the activation levels of a PDP network where there can be only one activation level per node. As long as differing regions of that Fourier space can have differential functional consequences, so long as they can exert differential control modulations, then the purely physical level of organization is split into multiple, perhaps a great many, interleaved but separable functional organizations. Differing regions of the Fourier space, in turn, can have differential functional *effects*, and, thus, constitute functionally different states, in terms of the Fourier space "region specificity" of the modulatory sensitivities of the modulat***ed*** processes. The possibility emerges of multiplexed functional processing. Still further, moving to such a Fourier space introduces intrinsic topologies into the space of system activities, and, therefore, into the interactively emergent system representations. It is only with respect to such topologies that the general problem of constructing heuristic learning and problem solving processes can be solved (Bickhard & Campbell, in preparation).

Binding

Note that if we shift perspective from the multiple Fourier space regions that might be *engaged in* differential control modulations to a

process that might be *being modulated* from multiple such processes — if we shift from the modulat*or* perspective to the modulat*ed* perspective — we find that oscillatory processes in differing regions of Fourier space and in differing physical parts of the system can exert joint modulatory influences over a single given process, as long as that given process is modulatorily sensitive to those multiple sources of influence. This is a natural solution to the "binding problem" within the oscillatory space model — the problem of how multiple representations that "go together" but that occur in differing parts of the overall system, such as the color and shape of an object, can be bound together for that system (Waltz & Feldman, 1988a; Sejnowski, 1986; Hummel & Biederman, 1992). In this architecture, binding is constituted by simultaneous modulatory control sensitivity of *subsequent* processes to the various bound processes.

Note that if the processes to be bound were all driven into similar regions of Fourier space, then subsequent binding sensitivity would be automatic, but that this would not be the only way to implement such binding. This model is roughly a generalization of the phase locking binding that has been explored — in which oscillations to be bound are driven into matching phases of oscillation (which requires matching frequencies) — but is much richer because of the richer potentialities of oscillatory spaces. It is also more natural in that modulation relationships will *intrinsically* tend to be differentially sensitive to differing regions of the oscillatory spaces, while phase locking requires, in general, some sort of phase modulatory driving to create it, and something exquisitely sensitive to phase — generally unspecified— to make use of, to detect and be sensitive to, the presence or absence of such phase locking. Note also that the Fourier space binding is a control relationship binding, not an encoded representation binding. Consequently, it is also more general in that sense, and can be functional for either control binding per se or for the representation binding emergent from such control binding.

It is interesting to note that phase binding is a special case of tensor binding in which the possible phase slots serve the tensor binding role (Shastri & Ajjanagadde, 1993; Smolensky, 1990; Tesar & Smolensky, 1994). The Fourier space version of binding could also be construed in such a manner, with the differential sensitivities to various regions of the Fourier space constituting the binding function. Because such sensitivities will, in general, be constructed along with the processes that manifest those sensitivities, the binding roles will not depend on predefined, and generally limited, binding slots. In this architecture,

being bound is constituted as sensitivities between some processes and others, while the *creation* of binding is the creation of binding sensitivities (potential slots), on the one hand, and the evocation of processes to be bound in the "right" regions of Fourier space.

Dynamic Trajectories

Standard PDP models involve regions of attraction for local attractors: it is the movement into one of the alternative attractors that is construed as generating the representation of the corresponding category of input patterns. The dynamic space of such a network, however, need not be organized strictly around local attractors. In particular, a network that has feedback of activation from later layers to earlier layers (note that with sufficient complexity, the division into layers becomes nugatory) can exhibit trajectory attractors rather than point or region attractors. In such a system, movement in the activation space will be attracted into one of alternative trajectories of further movement in the space. Such trajectory attractors can be closed, in which the activation vector travels around a closed loop, or open, in which the trajectory through the activation space does not cross itself, or branched, in which the branches might be contingent on subsequent inputs from the environment or elsewhere in the larger system, or chaotic, with their own interesting and emergent computational powers (Abraham & Shaw, 1992; Barnseley & Demko, 1986; Forrest, 1991; Garson, 1993; Haken, 1987; Pollack, 1991). For example, if the dynamic space of a system contains a chaotic region, transitions out of such a chaotic regime into dense spectrum of limit cycles — of attractors — yields a non-classically capturable capacity, even if any particular cycle might by itself be capturable by a classical rule: it would require an infinite number of rules for the attractors, and an infinite number of transition rules for the transitions into those attractors. If the spectrum of cycles were parameterizable, then perhaps a parameterized rule could capture it, but there is no apriori reason for it to be so.

Consider now an interactive system organization in which the dynamics of the system involved a loop trajectory in a dynamic space of oscillations. For a simple loop, such a system would constitute a higher level clock, with the cycles through the loop serving as the timing signal. If one particular part of that loop through the Fourier space (or n-fold product of Fourier spaces) served as a modulatory trigger for some other process, then it would be a timer for that process.

If the cycle were more complicated, however, perhaps shifting among several regions of the Fourier space, and if differing regions in that space served to modulate, to control, differing processes or aspects of processes elsewhere in the system, then the system would exert potentially complicated control over those other processes — and an intrinsically timed control. Such control trajectories could, in principle, be described as rules, but this would be *rule describable* control, not *rule governed* control: There are no rules anywhere in this system organization to be followed or obeyed.

If the trajectory of the controll*ing* system, in turn, were itself influenced by other modulatory influences from other parts of the system, the "rules" that would describe the overall functioning of the system would quickly become extremely complex. These could involve conditional shifts to differing trajectories, to differing loops, influences from the environment, recurrences under certain conditions, smooth manifolds of possible trajectories (loops or not) smoothly controlled by other processes, and so on. Conversely, the control influences from multiple such control trajectories could be bound, and bound differentially and even conditionally, in differing controll***ed*** processes. Any automaton or Turing machine organization could be manifested in the *intrinsic* dynamics of such a system — and, again, all with intrinsic timing, *and* intrinsic topology, making this architecture much more powerful than that of a Turing machine.

Such complex organization of control influences and resultant control dynamics is of potential interest in itself, and its simple equivalent in PDP networks with trajectories in their dynamics has been exploited (Dell, 1988; Elman, 1990, 1991; Rich & Knight, 1991), but the excitement has focused on input correspondences to the attractors, or more general points in the phase space, (mis)interpreted as representations of the input patterns that they differentiate.

There are exceptions, however. Exciting appreciations of the intrinsically dynamic character of thought can be found in the general "connectionist" literature (Clark, 1993; Smolensky, 1986, 1988; Churchland, 1986; Churchland & Sejnowski, 1992; van Gelder, 1992, in press-a). This extends, for example, to an appreciation both of the lack of timing in Turing machine theory and of the sense in which classical computational frameworks are special cases of the more general dynamic perspective. In effect, connectionism converges — partially — with the interactive perspective insofar as it converges with the dynamic systems

approach. There is as yet, however, no model of the emergence of representation within such a general dynamic approach — no model of interactive representation. Even when the importance of dynamics is appreciated, in fact, there is still an almost universal construal of representation in terms of correspondences between points in the phase space and things in the world — in terms of phase space points being interpreted as encodings (Smolensky, 1986, 1988; Churchland, 1986; Churchland & Sejnowski, 1992).

Complexly organized dynamic control relationships, of course, are of particular interest in the interactive framework. Representation is itself an emergent of precisely such complex control influences in the interactive view, so these kinds of possibilities are not simply of interest for control applications, but for the basic issues of representation and epistemology as well. Such an architecture should be sufficient for control processes beyond the capabilities of Turing machines, and for the emergence of representation for the system itself.

"Formal" Processes Recovered

A non-timed sequential control organization, should such be relevant (e.g., for formal "symbol manipulation" processes), could be constructed, for example, by the construction of conditional-test wait loops between each separate segment of the control process organization. Timing, but not sequence, would thus get abstracted out of the dynamics. This is exactly the case, for example, for symbol string rewrite rules, in which there is an inherent wait for matches of rewrite conditions to be found. Formal processes, then, are trivially recoverable within an interactive architecture; the reverse is impossible.

Differentiators In An Oscillatory Dynamics

The final states of an interactive differentiator have been discussed to this point only in terms of two or more nominal states in the set of possible final states. A further elaboration of the interactive architecture derives from noting that such final states will in general be themselves oscillatory processes, and, thus, that the set of possible such final states for a given differentiator will potentially have much more structure than a simple nominal set. In particular, in terms of the Fourier space, the set of final states will itself have a complex structure, perhaps an *n-fold* Fourier space. That differentiator space will in general be a multiple dimensional manifold, and the manifold will, in effect, parameterize the indicative or

modulatory influences of the differentiator on ensuing processes — such as the smooth manifolds of possible trajectories in a controlled process that was mentioned above. Still further, the modulatory relationships themselves are not necessarily all-or-nothing; they too could involve varying strengths, or more complex relationships, such as among various derivatives of the modulating and modulated processes.

An Alternative Mathematics

There is a rich mathematics for modeling and exploring such structures of dynamic spaces: dynamic systems theory and differential geometry (Abraham & Shaw, 1992; Burke, 1985; Hale & Koçak, 1991; Hermann, 1984; Nash & Sen, 1983; van Gelder & Port, in press; Ward & Wells, 1990; Warner, 1983; Wiggins, 1990).[23] For example, the dynamic spaces of the indicated ensuing processes will constitute fiber bundles over the manifolds of final states of differentiators, and the modulatory relationships will constitute various relationships among the tangent bundles of the relevant dynamic spaces. The richness and insights of such mathematics for understanding cognitive processes have yet to be explored.

The Interactive Alternative

The demands for rich intrinsic dynamic spaces and for intrinsic timing that interactivism imposes on any architecture of genuinely representational systems are not unrealizable. In fact, they yield the necessity of a kind of functional architecture very much like what is found in the brain — interactive architecture requires, and, in that sense, captures and explains, far more properties of neural functioning than does connectionist or PDP architectures.

Furthermore, interactive architecture provides power and possibilities that are not realizable within symbol manipulation or connectionist frameworks. Differing kinds of control influences constituted as differing forms of modulation, meta-modulation, etc.; the richness of oscillatory dynamic spaces; dynamic trajectories in those spaces as controllers; a natural basis for control — and, thus, representational — binding; and so on, are among the natural products of

[23] There are actually several different mathematical traditions involved, all approaching strongly overlapping areas of mathematics, but with differing notations, conceptualizations, and, in some cases, methods of proof. In the case of gauge theories in physics and differential geometry in mathematics, there was a significant historical lag before the mathematical commonalities were discovered (Atiyah, 1987).

the interactive architecture. And there is already a rich mathematics within which such dynamic phenomena can be investigated.

Most importantly, an architecture adequate to the intrinsic timing requirements of interactivism is necessary for any representational, intentional, system. In that sense, an interactive architecture is necessary for the ultimate programmatic aspirations of modeling, understanding, and constructing any representational, intentional, system. That is, an interactive architecture is necessary for the programmatic aspirations of Artificial Intelligence and Cognitive Science.

V

CONCLUSIONS

16

Transcending the Impasse

Artificial Intelligence and Cognitive Science are vast fields of research containing vital new developments — situated cognition, connectionism, language discourse analysis, to name but a few. Nevertheless, all of these are developments from within encodingism. Even the *rejections* of representation within some dynamic systems and robotics approaches are rejections of representations-as-encodings. There a number of partial insights into the problems of encodingism — though only rarely recognized as such — and many exciting partial moves away from it. But, they are only partial. As long as the fields remain within the encodingist framework, their basic programmatic aspirations cannot be met. Furthermore, at a less ambitious level, there are many vitiating intrusions of the incoherences and flaws of encodingism even into strictly practical research.

Artificial Intelligence and Cognitive Science contain their share, perhaps more than their share, of bluster, puffery, and scam. But these are not the basic problems. The most fundamental problem is the inheritance of a millennia old, false tradition of the nature of representation.

FAILURES OF ENCODINGISM

Encodingism fails the skepticism problem: all checking is circular. It fails against the copy argument: the system must already know what is to be copied — read, what is to have a representational correspondence set up for it — before it can construct the copy. It fails as a substance approach, both because it is a bad metaphysics, and because it is a metaphysics that cannot account for its own emergence — yielding wild claims of massive innateness of the basic representational atoms. Note that, although these conclusions of Fodor are widely dismissed because of their patent absurdity, they follow logically from the encodingism

assumption that underpins both fields: both fields are committed to them, or to something like them, whether they like it or not (Bickhard, 1991c).

Most fundamentally, encodingism fails on the incoherence problem. A benign God could, as for Descartes, ensure that our encodings are neither solipsistic illusions nor massively false. A foresighted God could make sure that all necessary encoding atoms were created along with the rest of the universe. A very busy God could watch over the individual constructions of instances of encodings to make sure that they are the right ones for the situations. But even God could not overcome the incoherence problem: to provide representational content in terms of some other encodings is to violate the assumption of base level atomicity, and to provide representational content in terms of some non-encoding representation is both to violate the assumption of basic level logical independence and to violate the encodingist framework altogether. To not provide representational content is to not address the problem.

Within encodingism, no primary intentionality, no "aboutness," is possible. The construction of new representational atoms is impossible, requiring that the original designed set be adequate to all future needs. Natural learning is impossible, both because of the impossibility of new representational atoms, and because of the absence of any natural error in systems without goal-directed interactive output. Encodingism is restricted to explicit representation. It is, therefore, incapable of capturing the unboundedness of implicit representation: of implicitly defined situations, predicated interactions, apperceptive updates, apperceptive contextualizations, and topologies of possible representations. In every case, encodingism faces an unbounded explosion of required explicit encodings.

With no intentionality, there can be no language understanding. So long as utterances are modeled as encodings rather than as operators — as interactions — language production *and* language understanding *and* the social processes of discourse will remain beyond reach. So too will the emergence of sociality and the co-constructed emergence of the social person (Bickhard, 1992a; in preparation-b).

Connectionism and PDP approaches fare no better in these regards. Trained correspondences are no improvement over transduced or designed correspondences with respect to basic epistemological issues. There is an irony in the fact that connectionism is supposed to model learning, yet the connectionist must design correspondences by training just as much as the classical symbol manipulationist must design them by

stipulation or engineering. (What does not have to be explicitly designed for PDP networks is the intrinsic topology in the dynamic space, and, thus, the intrinsic generalization.) Genuine learning in and from a natural environment will remain impossible without output — goal-directed, interactive output. Genuine learning requires interactivism.

Although we have not focused much on brain research in these discussions, note that, if encodingism is incoherent, then the brain does not and cannot function in accordance with the encodingist frameworks of GOFAI, connectionism, information processing, and so on. In adopting these standard frameworks, most contemporary neuroscience, too, is ensnared in a permeating encodingism (Bickhard, 1991e).

INTERACTIVISM

Interactivism is based upon interactive differentiation and its dual, implicit definition. Upon these are based functional indications of interactive possibilities. The task of maintaining such indications of interactive potentialities, of maintaining the situation image, is that of apperception. Perception looks quite different in this view than in standard encoding input models. The explicit situation image — the explicitly constructed indications of potential interaction — is a suborganization of the unbounded implicit situation image, implicitly defined by the constructive possibilities of the apperceptive procedures.

Because representation is constituted in interactive dynamic system organization, there is no possibility for representation to be merely impressed from the environment, either transductively or inductively. Learning and development require construction within the system itself, and, absent foreknowledge, this must be blind constructive variation with selective retention. Foreknowledge, of course, exists much, if not most, of the time, but its origin too must ultimately be accounted for (Bickhard, 1992a).

Language is a special form of interaction, distinguished by the special character of that which is being interacted with — social realities, situation conventions. Many properties of language and of utterances are constrained by the properties of situation conventions as objects of the interactions (Bickhard, 1980b) — all interactions must accommodate to that with which they interact. Neither utterance production nor utterance apperception are algorithmic, but can themselves pose problems that are addressable only with constructive variations and selections.

Interactivism forces an architecture with natural timing and natural topologies. This dynamic system architecture offers forms of control and computation more powerful than Turing machines. It also offers natural approaches to problems of binding across parallel processes. There is no message hang-up with the oscillatory modulations of this architecture. It also offers a natural framework within which representation can emerge, and, thus, in which learning and development can occur. In this view, it is no contingent accident of evolution that the brain functions in terms of modulations among oscillatory processes.

SOLUTIONS AND RESOURCES

In fact, interactivism offers an abundance of solutions to, or absences of, basic problems of encodingism. There is an intrinsic natural aboutness, intentionality, in functionally indicative predications. There are no empty symbols, no ungroundedness, no necessity for an interpreter, no circularity of representational content — no incoherence.

There is no aporia of emergence. Thus, no logically necessary but logically impossible innateness of representation. Interactive representation is emergently constructed as aspects of the construction of system organization itself.

Interactivism offers an intrinsic, natural, system-inherent form of error. There is no disjunction problem, no twin earth problem, no skepticism problem, no idealism problem. There is a natural ground of error for learning.

Interactive representation is intrinsically situated and embodied. Interactive differentiation and interactive predication are both intrinsically indexical and deictic and possible only in and for an embodied interactive system.

Interactive representation is intrinsically distributed over the spaces, the topologies, of functional indications. There is a very rich mathematics available for exploring such spaces. This is not just an implementational distributedness — to encapsulate and isolate any parts of such a space of modulations is to destroy the organization of relationships, of functional relationships, in which interactive representation is emergent. That is, interactive representation is intrinsically *relational*, functionally relational, and intrinsically *topological*, topologies of possible oscillations and modulations, and, thus, intrinsically and necessarily distributed. This necessity is not the case, for example, in connectionist nets.

Interactive representation is intrinsically implicit. Implicit definition is its grounding epistemic connection with the world. It is intrinsically modal — representation is of potential, possible, further interaction. Representation is of indications of modally possible interactions, and, thus, implicit predications of those implicitly defined interactive properties to the system environment. Interactive representation, therefore, is intrinsically unbounded. The spaces of implicitly defined possibilities are not bounded, and, therefore, cannot be exhaustively represented with explicit representations. Furthermore, the involvement of modalities here is itself intrinsic, not an ad hoc add-on to a non-modal logic (Bickhard, 1988a, 1988b; Bickhard & Campbell, 1989).

In fact, none of these properties of interactive representation are ad hoc. They all follow necessarily and intrinsically from the basic definition of interactive representation. They all come for free once the basic points of the encodingism critique and the interactive alternative are understood. They are all aporetic and impossible for *anything* within encodingism.

TRANSCENDING THE IMPASSE

Artificial Intelligence and Cognitive Science can accomplish many things from within the encoding approach, but they *cannot* accomplish *any* of their ultimate aspirations for understanding the mind and for creating artificial mentality from within that approach. Even at practical levels, the unboundedness of implicit representation and the consequent encodingist frame problems, the impossibility of capturing natural learning without the ability to construct new representations and to discover intrinsic error, and the impossibility of genuine language understanding without genuine representation, are examples of the fatal intrusions of encodingism.

Encodingism creates an intrinsic programmatic impasse. Encodingism distorts and intrudes its impossibilities and incoherencies into strictly practical efforts as well. Encodingism, if these considerations are at all valid, must be abandoned and transcended. Interactivism offers an alternative.

References

Abraham, R. H., Shaw, C. D. (1992). *Dynamics*. Addison-Wesley.

Acher, R. (1985). Evolution of Neuropeptides. In D. Bousfield (Ed.) *Neurotransmitters in Action.* (25-33) Amsterdam: Elsevier.

Adey, W. R. (1966). Neurophysiological Correlates of Information Transaction and Storage in Brain Tissue. In E. Stellar, J. M. Sprague (Eds.) *Progress in Physiological Psychology. Vol. 1*. (1-43) Academic.

Agnati, L. F., Fuxe, K., Pich, E. M., Zoli, M., Zini, I., Benfenati, F., Härfstrand, A., Goldstein, M. (1987). Aspects on the Integrative Capabilities of the Central Nervous System: Evidence for 'Volume Transmission' and its Possible Relevance for Receptor-Receptor Interactions. In K. Fuxe, L. F. Agnati (Eds.) *Receptor-Receptor Interactions*. (236-249) New York: Plenum.

Agre, P. E. (1988). The Dynamic Structure of Everyday Life. Ph.D. Dissertation, MIT AI Laboratory.

Agre, P. E., Chapman, D. (1987). Pengi: An Implementation of a Theory of Activity. *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, 196-201.

Allen, J. (1983). Recognizing Intentions from Natural Language Utterances. In M. Brady & R. Berwick (Ed.) *Computational Models of Discourse.* (107-166). Cambridge, MA: MIT Press.

Amarel, S. (1981). On Representations of Problems of Reasoning about Actions. In B. L. Webber, N. J. Nilsson (Eds.) *Readings in Artificial Intelligence*. (2-21). Los Altos, CA: Morgan Kaufmann.

Anderson, J. R. (1983). *The Architecture of Cognition.* Cambridge: Harvard University Press.

Angluin, D. (1992). Computational Learning Theory: Survey and Selected Bibliography. In *Proceedings of 24th Annual ACM Symposium on the Theory of Computing.* (351-368). Victoria, B.C., Canada. ACM.

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

Atiyah, M. (1987). *Michael Atiyah Collected Works. Vol. 5: Gauge Theories*. Oxford.

Atkinson, J. M., & Heritage, J. C. (1984). *Structures of Social Action: Studies in Conversation Analysis.* Cambridge: Cambridge University Press.

Austin, J. L. (1962). *How To Do Things With Words.* Cambridge: Harvard University Press.

Barnseley, M., Demko, S. (1986). *Chaotic Dynamics and Fractals*. Academic.

Barr, A., Cohen, P. R., Feigenbaum, E. A. (1981-1989). *The Handbook of Artificial Intelligence. Vols. I-IV*. Morgan Kaufmann; Addison-Wesley.

Barwise, J., Etchemendy, J. (1987). *The Liar*. Oxford.

Barwise, J., Etchemendy, J. (1989). Model-Theoretic Semantics. In M. I. Posner (Ed.) *Foundations of Cognitive Science.* (207-244). MIT.

Bechtel, W. (1986). Teleological functional analyses and the hierarchical organization of nature. In N. Rescher (Ed.) *Current Issues in Teleology.* (26-48). Landham, MD: University Press of America.

Bechtel, W. (1993). Currents in Connectionism. *Minds and Machines, 3*(2), 125-153.

Beer, R. D. (1990). *Intelligence as Adaptive Behavior*. Academic.

Beer, R. D. (in press-a). Computational and Dynamical Languages for Autonomous Agents. In R. Port, T. J. van Gelder (Eds.) *Mind as Motion: Dynamics, Behavior, and Cognition*. MIT.

Beer, R. D. (in press-b). A Dynamical Systems Perspective on Agent-Environment Interaction. *Artificial Intelligence*.

Beer, R. D., Chiel, H. J., Sterling, L. S. (1990). A Biological Perspective on Autonomous Agent Design. In P. Maes (Ed.) *Designing Autonomous Agents.* (169-186). MIT.

Beer, R. D., Gallagher, J. C. (1992). Evolving Dynamical Neural Networks for Adaptive Behavior. *Adaptive Behavior, 1*(1), 91-122.

Bickhard, M. H. (1973). A Model of Developmental and Psychological Processes. Ph. D. Dissertation, University of Chicago.

Bickhard, M. H. (1980a). A Model of Developmental and Psychological Processes. *Genetic Psychology Monographs, 102,* 61-116.

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

Bickhard, M. H. (1982). Automata Theory, Artificial Intelligence, and Genetic Epistemology. *Revue Internationale de Philosophie, 36,* 549-566.

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* (39-65). New York: Academic.

Bickhard, M. H. (1988a). Piaget on Variation and Selection Models: Structuralism, Logical Necessity, and Interactivism *Human Development, 31*, 274-312.

Bickhard, M. H. (1988b). The Necessity of Possibility and Necessity: Review of Piaget's Possibility and Necessity. *Harvard Educational Review, 58*, No. 4, 502-507.

Bickhard, M. H. (1991a). How to Build a Machine with Emergent Representational Content. *CogSci News*, *4*(1), 1-8.

Bickhard, M. H. (1991b). Homuncular Innatism is Incoherent: A reply to Jackendoff. *The Genetic Epistemologist, 19*(3), p. 5.

Bickhard, M. H. (1991c). The Import of Fodor's Anticonstructivist Arguments. In L. Steffe (Ed.) *Epistemological Foundations of Mathematical Experience*. (14-25). New York: Springer-Verlag.

Bickhard, M. H. (1991d). A Pre-Logical Model of Rationality. In L. Steffe (Ed.) *Epistemological Foundations of Mathematical Experience* (68-77). New York: Springer-Verlag.

Bickhard, M. H. (1991e). Cognitive Representation in the Brain. In *Encyclopedia of Human Biology*. *Vol. 2*. (547-558). Academic Press.

Bickhard, M. H. (1992a). How Does the Environment Affect the Person? In L. T. Winegar, J. Valsiner (Eds.) *Children's Development within Social Contexts: Metatheory and Theory*. (63-92). Erlbaum.

Bickhard, M. H. (1992b). Scaffolding and Self Scaffolding: Central Aspects of Development. In L. T. Winegar, J. Valsiner (Eds.) *Children's Development within Social Contexts: Research and Methodology*. (33-52). Erlbaum.

Bickhard, M. H. (1992c). Levels of Representationality. *Conference on The Science of Cognition*. Santa Fe, New Mexico, June 15-18.

Bickhard, M. H. (1992d). Myths of Science: Misconceptions of science in contemporary psychology. *Theory and Psychology, 2*(3), 321-337.

Bickhard, M. H. (1993a). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 285-333.

Bickhard, M. H. (1993b). On Why Constructivism Does Not Yield Relativism. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 275-284.

Bickhard, M. H. (1995). World Mirroring versus World Making: There's Gotta Be a Better Way. In L. Steffe (Ed.) *Constructivism in Education.* (229-267). Erlbaum.

Bickhard, M. H. (in preparation-a). From Epistemology to Rationality.

Bickhard, M. H. (in preparation-b). *The Whole Person: Toward a Naturalism of Persons.*

Bickhard, M. H. (in preparation-c). Interaction and Representation.

Bickhard, M. H. (in press-a). Intrinsic Constraints on Language: Grammar and Hermeneutics. *Journal of Pragmatics*.

Bickhard, M. H. (in press-b). Troubles with Computationalism. In R. Kitchener, W. O'Donohue (Eds.) *Psychology and Philosophy: Interdisciplinary Problems and Responses*.

Bickhard, M. H., & Campbell, D. T. (in preparation). Variations in Variation and Selection.

Bickhard, M. H., Campbell, R. L. (1989). Interactivism and Genetic Epistemology. *Archives de Psychologie, 57*(221), 99-121.

Bickhard, M. H., Campbell, R. L. (1992). Some Foundational Questions Concerning Language Studies: With a Focus on Categorial Grammars and Model Theoretic Possible Worlds Semantics. *Journal of Pragmatics, 17(5/6),* 401-433.

Bickhard, M. H., Campbell, R. L. (in preparation). Topologies of Learning and Development.

Bickhard, M. H., Christopher, J. C. (in press) The Influence of Early Experience on Personality Development. *New Ideas in Psychology*.

Bickhard, M. H., Cooper, R. G., Mace, P. E. (1985). Vestiges of Logical Positivism: Critiques of Stage Explanations. *Human Development, 28*, 240-258.

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

Bigelow, J., Pargetter, R. (1987). Functions. *Journal of Philosophy, 84*, 181-196.

Birkhoff, G. (1967). *Lattice Theory*. American Mathematical Society.

Bleicher, J. (1980). *Contemporary hermeneutics.* London: Routledge & Kegan Paul.

Block, N. (1980). Troubles with functionalism. In N. Block (Ed.) *Readings in philosophy and psychology* (Vol. I). (285-305). Cambridge: Harvard.

Block, N. (1986). Advertisement for a Semantics for Psychology. In P. A. French, T. E. Uehling, H. K. Wettstein (Eds.) *Midwest Studies in Philosophy X: Studies in the Philosophy of Mind*. (615-678) Minnesota.

Bloom, F. E., Lazerson, A. (1988). *Brain, Mind, and Behavior*. Freeman.

Bobrow, D. G. (1975). Dimensions of representation. In D. G. Bobrow, A. Collins (Eds.) *Representation and Understanding*. (1-34). New York: Academic.

Bobrow, D. G., Winograd, T. (1977). An Overview of KRL, a Knowledge Representation Language. *Cognitive Science 1*, 3-46.

Bobrow, D., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., & Winograd, T. (1977). GUS, A Frame-Driven Dialog System. *Artificial Intelligence 8*(2), 155-173.

Bogdan, R. (1988a). Information and Semantic Cognition: An Ontological Account. *Mind and Language, 3*(2), 81-122.

Bogdan, R. (1988b). Mental Attitudes and Common Sense Psychology. *Nous, 22*(3), 369-398.

Bogdan, R. (1989). What do we need concepts for? *Mind and Language, 4*(1,2), 17-23.

Bolinger, D. (1967). The Atomization of Meaning. In L. Jakobovits & M. Miron (Eds.) *Readings in the Psychology of Language*. (432-448). Englewood Cliffs, NJ: Prentice-Hall.

Boorse, C. (1976). Wright on Functions. *Philosophical Review, 85*, 70-86.

Booth, T. L. (1968). *Sequential Machines and Automata Theory.* New York: Wiley.

Bourgeois, P. L., Rosenthal, S. B. (1983). *Thematic Studies in Phenomenology and Pragmatism*. Amsterdam: Grüner.

Brachman, R. J. (1979). On the Epistemological Status of Semantic Networks. In N. V. Findler (Ed.) *Associative Networks: Representation and Use of Knowledge by Computers*. (3-50) New York: Academic.

Brainerd, W. S., Landweber, L. H. (1974). *Theory of Computation.* New York: Wiley.

Brooks, R. A. (1990). Elephants don't Play Chess. In P. Maes (Ed.) *Designing Autonomous Agents*. (3-15). MIT.

Brooks, R. A. (1991a). Intelligence without Representation. *Artificial Intelligence, 47*(1-3), 139-159.

Brooks, R. A. (1991b). New Approaches to Robotics. *Science, 253*(5025), 1227-1232.

Brooks, R. A. (1991c). How to Build Complete Creatures Rather than Isolated Cognitive Simulators. In K. VanLehn (Ed.) *Architectures for Intelligence*. (225-239). Erlbaum.

Brooks, R. A. (1991d). Challenges for Complete Creature Architectures. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (434-443). MIT.

Bullock, T. H. (1981). Spikeless Neurones: Where do we go from here? In A. Roberts & B. M. H. Bush (Eds.) *Neurones without Impulses*. (269-284). Cambridge University Press.

Burke, W. L. (1985). *Applied Differential Geometry*. Cambridge.

Burnyeat, M. (1983). *The Skeptical Tradition*. Berkeley: University of California Press.

Campbell, D. T. (1959). Methodological Suggestions from a Comparative Psychology of Knowledge Processes. *Inquiry, 2*, 152-182.

Campbell, D. T. (1974). Evolutionary Epistemology. In P. A. Schilpp (Ed.) *The Philosophy of Karl Popper*. (413-463). LaSalle, IL: Open Court.

Campbell, D. T. (1990). Levels of Organization, Downward Causation, and the Selection-Theory Approach to Evolutionary Epistemology. In Greenberg, G., & Tobach, E. (Eds.) *Theories of the Evolution of Knowing.* (1-17). Erlbaum.

Campbell, R. L. (1994). Could Categorial Perception Rescue Encodingism? Society for Philosophy and Psychology, Memphis, June 3.

Campbell, R. L., Bickhard, M. H. (1986). *Knowing Levels and Developmental Stages.* Basel: Karger.

Campbell, R. L., Bickhard, M. H. (1987). A Deconstruction of Fodor's Anticonstructivism. *Human Development, 30*(1), 48-59.

Campbell, R. L., Bickhard, M. H. (1992a). Clearing the Ground: Foundational Questions Once Again. *Journal of Pragmatics, 17(5/6),* 557-602.

Campbell, R. L., Bickhard, M. H. (1992b). Types of Constraints on Development: An Interactivist Approach. *Developmental Review, 12*(3), 311-338.

Carbonell, J. (1989). *Machine Learning.* MIT.

Changeux, J. (1991). Concluding Remarks. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (569-585) New York: Raven.

Chapman, D. (1987). Planning for Conjunctive Goals. *Artificial Intelligence, 32*(3), 333-377.

Chapman, D. (1991). *Vision, Instruction, and Action*. MIT.

Chapman, D., Agre, P. (1986). Abstract Reasoning as Emergent from Concrete Activity. In M. P. Georgeff, A. L. Lansky (Eds.) *Reasoning about Actions and Plans, Proceedings of the 1986 Workshop.* Timberline, Oregon, 411-424.

Chapman, M. (1988). *Constructive Evolution: Origins and Development of Piaget's Thought.* Cambridge: Cambridge University Press.

Cherian, S., Troxell, W. O. (1994a). A Distributed Control Strategy for Generating Useful Behavior in Animats. *Artificial Life IV*, July, MIT Press.

Cherian, S., Troxell, W. O. (1994b). Goal-Directedness in Behavior Networks. Workshop on Architectures for Intelligent Systems. Madrid, Spain, September 8-9, 1994.

Cherian, S., Troxell, W. O. (in press). Intelligent behavior in machines emerging from a collection of interactive control structures. *Computational Intelligence*, 1994. Blackwell Publishers. Cambridge, Mass. and Oxford, UK.

Chomsky, N. (1964). A Review of B. F. Skinner's *Verbal Behavior.* In J. A. Fodor, J. J. Katz (Eds.) *The Structure of Language.* (547-578). Prentice-Hall.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA:MIT Press.

Christiansen, M., Chater, N. (1992). Connectionism, Learning, and Meaning. *Connectionism, 4*, 227-252.

Churchland, P. M. (1989). *A Neurocomputational Perspective*. MIT.

Churchland, P. S. (1986). *Neurophilosophy*. MIT.

Churchland, P. S., Sejnowski, T. J. (1992). *The Computational Brain*. MIT.

Clancey, W. J. (1985). Heuristic Classification. *Artificial Intelligence, 27*, 289-350.

Clancey, W. J. (1989). The Knowledge Level Reinterpreted: Modeling How Systems Interact. *Machine Learning 4*, 285-291.

Clancey, W. J. (1991). The Frame of Reference Problem in the Design of Intelligent Machines. In K. VanLehn (Ed.) *Architectures for Intelligence: The Twenty-Second Carnegie Symposium on Cognition*. (357-423). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clancey, W. J. (1992a). The Knowledge Level Reinterpreted: Modeling Socio-Technical Systems. In Gaines, B. R. (Ed.) Working Notes of the AAAI Spring Symposium on *Cognitive Aspects of Knowledge Acquisition*.

Clancey, W. J. (1992b). Presentation at the AAAI Spring Symposium on *Cognitive Aspects of Knowledge Acquisition.*

Clancey, W. J. (1992c). Model Construction Operators. *Artificial Intelligence, 53*, 1-115.

Clancey, W. J. (1992d). Personal communication.

Clancey, W. J. (1993). Situated Action: A Neuropsychological Interpretation Response to Vera and Simon. *Cognitive Science, 17*(1), 87-116.

Clark, A. (1989). *Microcognition.* MIT.

Clark, A. (1993). *Associative Engines.* MIT.

Clark, A. (1994). Autonomous Agents and Real-Time Success: Some Foundational Issues. *PNP Research Report*, Washington University, St. Louis.

Clark, A., Toribio, J. (in press). Doing Without Representing? *Synthese*.

Cliff, D. (1991). Computational Neuroethology: A Provisional Manifesto. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (29-39). MIT.

Cliff, D. (1992). Neural Networks for Visual Tracking in an Artificial Fly. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (78-87). MIT.

Coffa, J. A. (1991). *The Semantic Tradition from Kant to Carnap.* Cambridge.

Cooper, J. R., Bloom, F. E., Roth, R. H. (1986). *The Biochemical Basis of Neuropharmacology*. Oxford.

Cowan, J. D., Sharp, D. H. (1988). Neural Nets and Artificial Intelligence. In S. R. Graubard (Ed.) *The Artificial Intelligence Debate*. (85-121) MIT.

Craig, W. (1974). *Logic in Algebraic Form*. Amsterdam: North-Holland.

Cummins, R. (1975). Functional Analysis. *Journal of Philosophy, 72*, 741-764.

Cussins, A. (1990). The Connectionist Construction of Concepts. In M. A. Boden (Ed.) *The Philosophy of Artificial Intelligence*. (368-440). Oxford.

Cussins, A. (1992). The Limitations of Pluralism. In D. Charles, K. Lennon (Eds.) *Reduction, Explanation, and Realism*. (179-223). Oxford.

Cutland, N. J. (1980). *Computability*. Cambridge.

Dell, G. S. (1988). Positive Feedback in Hierarchical Connectionist Models. In D. Waltz, J. A. Feldman (Eds.) *Connectionist Models and Their Implications*. (97-117). Norwood, NJ: Ablex.

Demopoulos, W., Friedman, M. (1989). The Concept of Structure in *The Analysis of Matter*. In C. W. Savage, C. A. Anderson (Eds.) *Rereading Russell*. (183-199). U. of Minnesota.

Dennett, D. C. (1987). *The Intentional Stance*. MIT.

Dennett, D. C. (in press). Producing Future by Telling Stories. In K. M. Ford, Z. Pylyshyn (Eds.) *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Ablex.

Dowling, J. E. (1992). *Neurons and Networks*. Harvard.

Drescher, G. L. (1986). Genetic AI: Translating Piaget into Lisp. MIT AI Memo No. 890.

Drescher, G. L. (1991). *Made-Up Minds*. MIT.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. MIT.

Dretske, F. I. (1988). *Explaining Behavior*. MIT.

Dreyfus, H. L. (1967). Why Computers Must Have Bodies in order to be Intelligent. *Review of Metaphysics, 21*, 13-32.

Dreyfus, H. L. (1979). *What Computers Can't Do. 2nd ed.* New York: Harper & Row.

Dreyfus, H. L. (1981). From micro-worlds to knowledge representation: AI at an impasse. In J. Haugeland (Ed.) *Mind design*. (161-204). Cambridge: MIT. Press.

Dreyfus, H. L. (1982). Introduction. In H. L. Dreyfus (Ed.) *Husserl: Intentionality & Cognitive Science*. (1-27). MIT.

Dreyfus, H. L. (1991). *Being-in-the-World.* MIT.

Dreyfus, H. L. (1992). *What Computers **Still** Can't Do*. MIT.

Dreyfus, H. L., Dreyfus, S. E. (1986). *Mind Over Machine*. New York: Free Press.

Dreyfus, H. L., Dreyfus, S. E. (1987). How to Stop Worrying about the Frame Problem even though its Computationally Insoluble. In Z. W. Pylyshyn (Ed.) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence.* (95-111). Norwood, NJ: Ablex.

Dreyfus, H. L., Dreyfus, S. E. (1988). Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint. In S. R. Graubard (Ed.) *The Artificial Intelligence Debate.* (15-43). MIT.

Dreyfus, H. L., Haugeland, J. (1978). Husserl and Heidegger: Philosophy's Last Stand. In M. Murray (Ed.) *Heidegger & Modern Philosophy* (222-238). Yale University Press.

Dummett, M. (1973). *Frege: Philosophy of language.* New York: Harper & Row.

Dunnett, S. B., Björklund, A., Stenevi, U. (1985). Dopamine-rich Transplants in Experimental Parkinsonism. In D. Bousfield (Ed.) *Neurotransmitters in Action*. (261-269) Amsterdam: Elsevier.

Eco, U., Santambrogio, M., Violi, P. (1988). *Meaning and Mental Representations*. Indiana University Press.

Eilenberg, S. (1974). *Automata, Languages, and Machines. Vol. A.* Academic.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-212.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-225.

Emson, P. C. (1985). Neurotransmitter Systems. In D. Bousfield (Ed.) *Neurotransmitters in Action.* (6-10) Amsterdam: Elsevier.

Field, H. (1980). Tarski's Theory of Truth. In M. Platts (Ed.) *Reference, Truth and Reality.* (83-110). London: Routledge.

Field, H. (1981). Mental Representation. In N. Block (Ed.) *Readings in philosophy and psychology* (Vol. II). (78-114). Cambridge: Harvard.

Fikes, R., & Nilsson, N. (1971). STRIPS: A New Approach to the Application of Theorem Proving in Problem Solving. *Artificial Intelligence 2*(3-4), 189-208.

Findler, N. V. (1979). *Associate Networks: The Representation and Use of Knowledge by Computers*. New York: Academic Press.

Fischer, G. (1990). Communication Requirements for Cooperative Problem Solving Systems. *International Journal of Intelligent Systems* (*Special Issue on Knowledge Engineering), 15*(1), 21-36.

Fodor, J. A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. A. (1978). Tom Swift and his Procedural Grandmother. *Cognition 6*, 229-247.

Fodor, J. A. (1981a). Methodological solipsism considered as a research strategy in cognitive psychology. In J. Haugeland (Ed.) *Mind design*. (307-338). Cambridge: MIT. Press.

Fodor, J. A. (1981b). The present status of the innateness controversy. In J. Fodor *RePresentations* (257-316). Cambridge: MIT Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* MIT.

Fodor, J. A. (1987). *Psychosemantics.* MIT.

Fodor, J. A. (1990). *A Theory of Content and Other Essays.* MIT.

Fodor, J. A. (1990b). Information and Representation. In P. P. Hanson (Ed.) *Information, Language, and Cognition.* (175-190). University of British Columbia Press.

Fodor, J. A. (1994). *The Elm and the Expert*. MIT.

Fodor, J. A., Bever, T., Garrett, M. (1974). *The Psychology of Language.* New York: McGraw-Hill.

Fodor, J. A., Pylyshyn, Z. (1981). How direct is visual perception?: Some reflections on Gibson's ecological approach. *Cognition, 9*, 139-196.

Fodor, J. A., Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. In S. Pinker, J. Mehler (Eds.) *Connections and Symbols*. (3-71). Cambridge: MIT.

Ford, K. M. (1989). A constructivist view of the frame problem in artificial intelligence. *Canadian Psychology, 30*, 188-190.

Ford, K. M., Adams-Webber, J. R. (1992). Knowledge acquisition and constructivist epistemology. In R. R. Hoffman (Ed.) *The Psychology of Expertise: Empirical Approaches to Knowledge Acquisition*. (121-136) New York: Springer-Verlag.

Ford, K. M., Agnew, N. M. (1992). Expertise: Socially situated and personally constructed. In *Working Notes of the Cognitive Aspects of Knowledge Acquisition Session of the AAAI Spring Symposium,* Stanford, 80-87.

Ford, K. M., Agnew, N., Adams-Webber, J. R. (in press). Goldilocks and the Frame Problem. In K. M. Ford & Z. Pylyshyn (Eds.) *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex Press.

Ford, K. M., Hayes, P. J. (1991). *Reasoning Agents in a Dynamic World: The Frame Problem.* Greenwich, CT: JAI Press.

Forrest, S. (1991). *Emergent Computation*. MIT.

Freeman, W. J., Skarda, C. A. (1990). Representations: Who Needs Them? In J. L. McGaugh, N. M. Weinberger, G. Lynch (Eds.) *Brain Organization and Memory.* (375-380) Oxford.

Fuster, J. M. (1989). *The Prefrontal Cortex*. New York: Raven Press.

Fuxe, K., Agnati, L. F. (1987). *Receptor-Receptor Interactions.* New York: Plenum.

Fuxe, K., Agnati, L. F. (1991a). *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission.* New York: Raven.

Fuxe, K., Agnati, L. F. (1991b). Two Principal Modes of Electrochemical Communication in the Brain: Volume versus Wiring Transmission. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (1-9) New York: Raven.

Gadamer, Hans-Georg (1975). *Truth and method*. New York: Continuum.

Gadamer, Hans-Georg (1976). *Philosophical hermeneutics*. Berkeley: University of California Press.

Gallagher, J. C., Beer, R. D. (1993). A Qualitative Dynamical Analysis of Evolved Locomotion Controllers. In J.-A. Meyer, H. L. Roitblat, S. W. Wilson (Eds.) *From Animals to Animats 2*. (71-80). MIT.

Gallistel, C. R. (1980). *The Organization of Action: A New Synthesis*. Hillsdale, NJ: Lawrence Erlbaum.

Gallistel, C. R. (1990). *The Organization of Learning.* MIT.

Garfinkel, H. (1967a). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garfinkel, H. (1967b). What is Ethnomethodology? In H. Garfinkel *Studies in Ethnomethodology*. (1-34). Englewood Cliffs, NJ: Prentice-Hall.

Garson, J. W. (1993). Chaotic Cognition: Prospects for Symbolic Processing in a Dynamic Mind. Society for Philosophy and Psychology, Vancouver, B.C., June 1, 1993.

Genesereth, M. R., Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence.* Morgan Kaufmann.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.) *Perceiving, acting and knowing*. (67-82). Hillsdale, NJ: Erlbaum.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.) *Unintended Thought.* (189-211). New York: Guilford.

Ginzburg, A. (1968). *Algebraic Theory of Automata*. Academic.

Glass, A. L., Holyoak, K. J., Santa, J. L. (1979). *Cognition*. Reading, MA: Addison-Wesley.

Glymour, C. (1987). Android Epistemology and the Frame Problem. In Z. Pylyshyn (Ed.) *The Robot's Dilemma*. (65-76). Ablex.

Goschke, T., Koppelberg, D. (1991). The Concept of Representation and the Representation of Concepts in Connectionist Models.. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.) *Philosophy and Connectionist Theory*. (129-161). Erlbaum.

Grandy, R. (1979). *Advanced logic for applications*. Dordrecht, Holland: Reidel.

Greenbaum, J., Kyng, M. (Eds.) (1991). *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.) *Syntax and semantics. Vol. III: Speech acts*. (41-58). New York: Academic Press.

Groarke, L. (1990). *Greek Scepticism*. Montreal: McGill-Queens.

Guha, R., Lenat, D. (1988). CycLing: Inferencing in Cyc. MCC Technical Report No. ACA-AI-303-88.

Guignon, C. B. (1983). *Heidegger and the Problem of Knowledge.* Indianapolis: Hackett.

Gupta, A., Belnap, N. (1993). *The Revision Theory of Truth*. MIT.

Habermas, J. (1971). *Knowledge and Human Interests.* Boston: Beacon.

Habermas, J. (1979). *Communication and the Evolution of Society*. Boston: Beacon.

Hadley, R. F. (1989). A Default-Oriented Theory of Procedural Semantics. *Cognitive Science, 13*(1), 107-137.

Haken, H. (1987). *Computational Systems — Natural and Artificial.* Springer-Verlag.

Hale, J. K., Koçak, H. (1991). *Dynamics and Bifurcations.* Springer-Verlag.

Hall, Z. W. (1992). *Molecular Neurobiology*. Sunderland, MA: Sinauer.

Hanson, P. P. (1990). *Information, Language, and Cognition*. University of British Columbia Press.

Hansson, E. (1991). Transmitter Receptors on Astroglial Cells. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (257-265) New York: Raven.

Harman, G. (1982). Conceptual Role Semantics. *Notre Dame Journal of Formal Logic, 23*(2), 242-256.

Harman, G. (1987). (Nonsolipsistic) Conceptual Role Semantics. In E. LePore (Ed.) *New Directions in Semantics*. (55-81). Academic.

Harnad, S. (1987a). Psychophysical and Cognitive Aspects of Categorical Perception: A Critical Overview. In S. Harnad (Ed.) *Categorial Perception: The Groundwork of Cognition.* (1-25) NY: Cambridge.

Harnad, S. (1987b). Category Induction and Representation. In S. Harnad *Categorial Perception: The Groundwork of Cognition*. (535-565). New York: Cambridge.

Harnad, S. (1989). Minds, Machines and Searle. *JETAI, 1*, 5-25.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D, 42*, 335-346.

Harnad, S. (1991). Other Bodies, Other Minds: A machine incarnation of an old philosophical problem. *Minds and Machines, 1*, 43-54.

Harnad, S. (1992a). Connecting Object to Symbol in Modeling Cognition. In A. Clark, R. Lutz (Eds.) *Connectionism in Context*. (75-90). Springer-Verlag.

Harnad, S. (1992b). The Turing Test Is Not A Trick: Turing Indistinguishability is a Scientific Criterion. *SIGART Bulletin 3*(4) (Oct 92) 9-10.

Harnad, S. (1993a). Grounding Symbols in the Analog World with Neural Nets: A Hybrid Model. *Think*, 2, whole issue.

Harnad, S. (1993b). Artificial Life: Synthetic versus Virtual. *Artificial Life III*, Santa Fe, June 1992.

Harnad, S. (1993c). The Origin of Words: A Psychophysical Hypothesis. In W. Durham, B. Velichkovsky *Proceedings of the Zif Conference on Biological and Cultural Aspects of Language Development*. Muenster: Nodus Publishers.

Harnad, S. (1993d). Symbol Grounding is an Empirical Problem: Neural Nets are Just a Candidate Component. *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society.* Erlbaum.

Harnad, S., Hanson, S. J., Lubin, J. (1991). Categorial Perception and the Evolution of Supervised Learning in Neural Nets. In D. W. Powers, L. Reeker (Eds.) *Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*. (65-74).

Harnad, S., Hanson, S. J., Lubin, J. (1994). Learned Categorical Perception in Neural Nets: Implications for Symbol Grounding. In V. Honavar & L. Uhr (Eds.) *Symbol Processing and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration.* (191-206). Academic.

Hartmanis, J., Stearns, R. E. (1966). *Algebraic Structure Theory of Sequential Machines*. Englewood Cliffs, NJ: Prentice-Hall.

Hatfield, G. (1986). Representation and Content in Some (Actual) Theories of Perception. Baltimore, Maryland: Johns Hopkins University, Reports of the Cognitive Neuropsychology Laboratory #21.

Hatfield, G. (1987). Information and Representation in Noncognitive Explanations of Perception. Manuscript.

Haugeland, J. (1985). *Artificial Intelligence*. MIT.

Haugeland, J. (1991). Representational Genera.. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.) *Philosophy and Connectionist Theory.* (61-89). Erlbaum.

Hayes, P. J., & Ford, K. M. (in preparation). Closing the Door on the Chinese Room.

Hayes, P. J., Ford, K. M., Adams-Webber, J. R. (1992). Human Reasoning about Artificial Intelligence. *Journal of Experimental and Theoretical Artificial Intelligence, 4*, 247-263.

Hayes, P., Harnad, S., Perlis, D., & Block, N. (1992). Virtual Symposium on Virtual Mind. *Minds and Machines, 2*(3), 217-238.

Heidegger, M. (1962). *Being and Time.* New York: Harper & Row.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

Hendler, J., Tate, A., & Drummond, M. (1990). AI Planning: Systems and Techniques. *AI Magazine 11*(2), 61-77.

Henkin, L., Monk, J., & Tarski, A. (1971). *Cylindric Algebras*. Amsterdam: North Holland.

Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Cambridge, England: Polity Press.

Herken, R. (1988). *The Universal Turing Machine*. Oxford: Oxford University Press.

Herkenham, M. (1991). Mismatches Between Neurotransmitter and Receptor Localizations: Implications for Endocrine Functions in Brain. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (63-87) New York: Raven.

Hermann, R. (1984). *Topics in the Geometric Theory of Linear Systems*. Brookline, MA: Math Sci Press.

Herstein, I. N. (1964). *Topics in Algebra*. New York: Blaisdell.

Herzberger, H. G. (1970). Paradoxes of Grounding in Semantics. *Journal of Philosophy, 67*, 145-167.

Hille, B. (1987). Evolutionary Origins of Voltage-Gated Channels and Synaptic Transmission. In G. M. Edelman, W. E. Gall, W. M. Cowan (Eds.) *Synaptic Function*. (163-176). New York: Wiley.

Hodges, A. (1983). *Alan Turing: The Enigma*. New York: Simon and Schuster.

Hodges, A. (1988). Alan Turing and the Turing Machine. In R. Herken (Ed.) *The Universal Turing Machine*. (3-15). Oxford: Oxford University Press.

Hollan, J. D., Miller, J. R., Rich, E., & Wilner, W. (1991). Knowledge Bases and Tools for Building Integrated Multimedia Intelligent Interfaces. In Sullivan, J. W. & Tyler, S. W. (Ed.) *Intelligent User Interfaces*. (293-337). Reading, MA: Addison-Wesley.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, R. T. (1986) *Induction*. Cambridge: MIT Press.

Honavar, V. (in press). Symbolic Artificial Intelligence and Numeric Artificial Neural Networks: Towards a Resolution of the Dichotomy. In: Sun, R. and Bookman, L. (Ed.) *Computational Architectures Integrating Symbolic and Neural Processes*. New York: Kluwer.

Honavar, V., Uhr, L. (1994). *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*. San Diego, CA: Academic.

Hooker, C. A. (1994, in press). *Reason, Regulation, and Realism: Towards a Regulatory Systems Theory of Reason and Evolutionary Epistemology*. SUNY.

Hooker, C. A. (in preparation). Toward a Naturalised Cognitive Science.

Hooker, C. A., Penfold, H. B., Evans, R. J. (1992). Towards a Theory of Cognition Under a New Control Paradigm. *Topoi, 11*, 71-88.

Hoopes, J. (1991). *Peirce on Signs*. Chapel Hill.

Hopcroft, J. E., Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley.

Horgan, T. (1993). From Supervenience to Superdupervenience: Meeting the Demands of a Material World. *Mind, 102*(408), 555-586.

Horgan, T., Tienson, J. (1988). Settling into a New Paradigm. The Spindel Conference 1987: Connectionism and the Philosophy of Mind. *The Southern Journal of Philosophy, XXVI*(supplement), 97-114.

Horgan, T., Tienson, J. (1992). Cognitive Systems as Dynamical Systems. *Topoi, 11*, 27-43.

Horgan, T., Tienson, J. (1993). Levels of Description in Nonclassical Cognitive Science. *Philosophy, 34, Supplement*, 159-188.

Horgan, T., Tienson, J. (1994). A Nonclassical Framework for Cognitive Science. Manuscript.

Houser, N., Kloesel, C. (1992). *The Essential Peirce. Vol. 1*. Indiana.

Howard, R. J. (1982). *Three faces of hermeneutics.* Berkeley: U. of California Press.

Hummel, J. E., Biederman, I. (1992). Dynamic Binding in a Neural Network for Shape Recognition. *Psychological Review, 99*(3), 480-517.

Husain, M. (1983). To *What* can One Apply a Construct? In J. R. Adams-Webber, J. C. Mancuso (Eds.) *Applications of Personal Construct Psychology*. (11-28) New York: Academic Press.

Iverson, L. L., Goodman, E. (1986). *Fast and Slow Chemical Signalling in the Nervous System.* Oxford.

Johnson-Laird, P. (1977). Procedural Semantics. *Cognition, 5*, 189-214.

Johnson-Laird, P. (1978). What's Wrong with Grandma's Guide to Procedural Semantics. *Cognition, 6*, 262-271.

Johnson-Laird, P. N. (1983). *Mental Models.* Cambridge: Harvard.

Jordan, M. J., Rumelhart, D. E. (1992). Forward Models: Supervised Learning with a Distal Teacher. *Cognitive Science, 16*, 307-354.

Kaelbling, L. P. (1986). An architecture for intelligent reactive systems. In M. P. Georgeff, A. L. Lansky (Eds.) *Proceedings of the 1986 Workshop on Reasoning about Actions and Plans*. (395-410). Los Altos, CA: Morgan Kaufmann.

Kaelbling, L. P. (1992). An Adaptable Mobile Robot. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (41-47). MIT.

Kalat, J. W. (1984). *Biological Psychology. 2nd Edition*. Belmont, CA: Wadsworth.

Kandel, E. R., Schwartz, J. H. (1985). *Principles of Neural Science. 2nd ed*. New York: Elsevier.

Kaplan, D. (1979a). On the logic of demonstratives. In P. French, T. Uehling, Jr., & H. Wettstein (Eds.) *Contemporary Perspectives in the Philosophy of Language*. (401-412). Minneapolis: U. of Minnesota Press.

Kaplan, D. (1979b). Dthat. In P. French, T. Uehling, Jr., & H. Wettstein (Eds.) *Contemporary Perspectives in the Philosophy of Language*. (383-400). Minneapolis: U. of Minnesota Press.

Kaplan, D. (1979c). Transworld Heir Lines. In M. J. Loux (Ed.) *The Possible and the Actual*. (88-109) Cornell.

Kaplan, D. (1989). Demonstratives: an essay on semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Allmog, J. Perry, H. Wettstein (Eds.) *Themes from Kaplan*. (481-563). Oxford University Press.

Katz, J. J., Fodor, J. A. (1971). The structure of a semantic theory. In J. F. Rosenberg & C. Travis (Eds.) *Readings in the Philosophy of Language*. (472-514). Englewood Cliffs, NJ: Prentice-Hall.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Kitchener, R. F. (1986). *Piaget's Theory of Knowledge*. New Haven: Yale.

Koch, C., Poggio, T. (1987). Biophysics of Computation. In G. M. Edelman, W. E. Gall, W. M. Cowan (Eds.) *Synaptic Function*. (637-697). New York: Wiley.

Korf, R. E. (1985). Macro-operators: A weak method for learning. *Artificial Intelligence, 26*, 35-77.

Kosslyn, S. M., Hatfield, G. (1984). Representation without Symbol Systems. *Social Research, 51*(4), 1019-1045.

Krall, P. (1992). A Model for Procedural Representation as a Basis for Adaptive Self-modification. *Evolution and Cognition, 2*, 211-231.

Kripke, S. A. (1972). Naming and Necessity. In D. Davidson, G. Harman (Eds.) *Semantics of Natural Language*. (253-355). Reidel.

Kuipers, B. J. (1988). The TOUR Model: A Theoretical Definition. From Kuipers, B. J., Levitt, T. Navigation and Mapping in Large-Scale Space. *AI Magazine, 9(*2), 25-43.

Kuipers, B. J., Byun, Y. (1991). A Robot Exploration and Mapping Strategy Based on a Semantic Hierarchy of Spatial Representations. *Robotics and Autonomous Systems, 9*, 47-63.

Kyburg, H. E. (in press). Dennett's Beer. In K. M. Ford, Z. Pylyshyn (Eds.) *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Ablex.

Lai, K. Y., Malone, T. W., & Yu, K. C. (1988). Object Lens: A "Spreadsheet" for Cooperative Work. *ACM Transactions on Office Information Systems, 6*, 332-353.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An Architecture for General Intelligence. *Artificial Intelligence, 33*, 1-64.

Laird, J. E., Rosenbloom, P. S., Newell, A. (1984). Toward Chunking as a General Learning Mechanism. *Proceedings of the National Conference on Artificial Intelligence*, Austin, TX, 188-192.

Laird, J. E., Rosenbloom, P. S., Newell, A. (1986). Chunking in SOAR: The Anatomy of a General Learning Mechanism. *Machine Learning, 1*(1), Jan 86, 11-46.

Lakoff, G., Johnson, M. (1980). *Metaphors We Live By.* Chicago.

Lapedes, A., Farber, R. (1986). A Self-Optimizing, Nonsymmetrical Neural Net for Content Addressable Memory and Pattern Recognition. *Physica D, 22*, 247.

Lave, J. (1988). *Cognition in Practice*. Cambridge: Cambridge University Press.

Lenat, D. B., Feigenbaum, E. A. (1991). On the Thresholds of Knowledge. *Artificial Intelligence, 47*(1-3), 185-250.

Lenat, D., Guha, R. (1988). The World According to CYC. MCC Technical Report No. ACA-AI-300-88.

Lenat, D., Guha, R., Pittman, K., Pratt, D., Shephard, M. (1990). CYC: Toward Programs with Common Sense. *Communications of the ACM, 33*(8), 30-49.

Lenat, D., Guha, R., Wallace, D. (1988). The CycL Representation Language. MCC Technical Report No. ACA-AI-302-88.

Levesque, H. J. (1988). Logic and the complexity of reasoning. *Journal of Philosophical Logic, 17*(4), 355-389.

Levinson, S. C. (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Litman, D. (1985). Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues. Technical Report No. 170 and Ph.D. Thesis. Rochester, NY: The University of Rochester.

Litman, D., Allen, J. (1987). A Plan Recognition Model for Subdialogues in Conversations. *Cognitive Science* 11, 163-200.

Loewer, B., Rey, G. (1991). *Meaning in Mind: Fodor and his critics.* Blackwell.

Loux, M. J. (1970). *Universals and Particulars*. Notre Dame.

Luff, P., Gilbert, N., Frohlich, D. (1990). *Computers and Conversation*. London: Academic Press.

MacKay, D. G. (1987). *The Organization of Perception and Action*. New York: Springer-Verlag.

MacLane, S. (1971). *Categories for the Working Mathematician*. New York: Springer-Verlag.

MacLane, S. (1986). *Mathematics: Form and Function*. New York: Springer-Verlag.

MacLane, S., Birkhoff, G. (1967). *Algebra*. New York: Macmillan.

Maes, P. (1990a). *Designing Autonomous Agents.* MIT.

Maes, P. (1990b). Situated Agents Can Have Goals. In P. Maes (Ed.) *Designing Autonomous Agents*. (49-70). MIT.

Maes, P. (1991). A Bottom-Up Mechanism for Behavior Selection in an Artificial Creature. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats.* (238-246). MIT.

Maes, P. (1992). Learning Behavior Networks from Experience. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (48-57). MIT.

Maes, P. (1993). Behavior-Based Artificial Intelligence. In J.-A. Meyer, H. L. Roitblat, S. W. Wilson (Eds.) *From Animals to Animats 2*. (2-10). MIT.

Maes, P. (1994). Modeling Adaptive Autonomous Agents. *Artificial Life, 1*, 135-162.

Malcolm, C. A., Smithers, T., Hallam, J. (1989). An Emerging Paradigm in Robot Architecture. In T Kanade, F.C.A. Groen, & L.O. Hertzberger (Eds.) *Proceedings of the Second Intelligent Autonomous Systems Conference*. (284-293). Amsterdam, 11--14 December 1989. Published by Stichting International Congress of Intelligent Autonomous Systems.

Malcolm, C., Smithers, T. (1990). Symbol Grounding via a Hybrid Architecture in an Autonomous Assembly System. In P. Maes (Ed.) *Designing Autonomous Agents.* (123-144). MIT.

Manfredi, P. A. (1986). Processing or pickup: Conflicting approaches to perception. *Mind & Language*, 1(3), 181-200.

Martin, R. L. (1984). *Recent Essays on Truth and the Liar Paradox*. Oxford.

Mataric, M. J. (1991). Navigating With a Rat Brain: A Neurobiologically-Inspired Model for Robot Spatial Representation. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (169-175). MIT.

Matteoli, M., Reetz, A. T., De Camilli, P. (1991). Small Synaptic Vesicles and Large Dense-Core Vesicles: Secretory Organelles Involved in Two Modes of Neuronal Signaling. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (181-193) New York: Raven.

Maturana, H. R., Varela, F. J. (1980). *Autopoiesis and cognition.* Dordrecht, Holland: Reidel.

Maturana, H. R., Varela, F. J. (1987). *The Tree of Knowledge*. Boston: New Science Library.

McCarthy, J., Hayes, P. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. In B. Meltzer, D. Michie (Eds.) *Machine Intelligence 4*. (463-502). New York: American Elsevier.

McClelland, J. L., Rumelhart, D. E. (1986). *Parallel Distributed Processing. Vol. 2: Psychological and Biological Models.* MIT.

McDermott, D. (1981). Artificial Intelligence meets Natural Stupidity. In J. Haugeland (Ed.) *Mind Design*. (143-160). MIT.

Mehan, H., Wood, H. (1975). *The Reality of Ethnomethodology*. New York: Wiley.

Melton, A. W., Martin, E. (1972). *Coding Processes in Human Memory.* New York: Wiley.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. Henry Holt and Company.

Miller, G. A., Johnson-Laird, P. N. (1976). *Language and Perception*. Cambridge: Harvard.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. MIT.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. MIT.

Minsky, M. (1967). *Computation.* Englewood Cliffs, NJ.: Prentice-Hall.

Minsky, M. (1981). A Framework for Representing Knowledge. In J. Haugeland (Ed.) *Mind design*. (95-128). Cambridge: MIT. Press.

Minsky, M., Papert, S. (1969). *Perceptrons*. MIT.

Murphy, J. P. (1990). *Pragmatism*. Westview.

Nash, C., Sen, S. (1983). *Topology and Geometry for Physicists*. Academic.

Nauta, W. J. H., Feirtag, M. (1986). *Fundamental Neuroanatomy.* Freeman.

Neander, K. (1991). Functions as Selected Effects: The Conceptual Analyst's Defense. *Philosophy of Science, 58*(2), 168-184.

Neches, R., Langley, P., & Klahr, D. (1987). Learning, development, and production systems. In D. Klahr, P. Langley, & R. Neches (Eds.) *Production system models of learning and development* (1-53). Cambridge, MA: MIT Press.

Nedergaard, M. (1994). Direct Signaling from Astrocytes to Neurons in Cultures of Mammalian Brain Cells. *Science, 263*, 1768-1771.

Nehmzow, U., Smithers, T. (1991). Mapbuilding Using Self-Organizing Networks in "Really Useful Robots." In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (152-159). MIT.

Nehmzow, U., Smithers, T. (1992). Using Motor Actions for Location Recognition. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (96-104). MIT.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Newell, A. (1980a). Physical Symbol Systems. *Cognitive Science, 4*, 135-183.

Newell, A. (1980b). Reasoning, Problem Solving, and Decision Processes: The Problem Space as a Fundamental Category. In R. Nickerson (Ed.) *Attention and Performance VIII*. (693-718). Hillsdale, NJ: Erlbaum.

Newell, A. (1982). The Knowledge Level. *Artificial Intelligence 18*(1), 87-127.

Newell, A., Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., Simon, H. A. (1975). Computer Science as Empirical Inquiry: Symbols and Search. In (1987). *ACM Turing Award Lectures: The First Twenty Years, 1966-1985*. (287-313). New York: ACM Press; Reading, Mass.: Addison-Wesley Pub. Co.

Newell, A., Simon, H. A. (1987). Postscript: Reflections on the Tenth Turing Award Lecture: Computer Science as Empirical Inquiry — Symbols and Search. In *ACM Turing Award Lectures: The First Twenty Years, 1966-1985*. (314-317). New York: ACM Press; Reading, Mass.: Addison-Wesley Pub. Co.

Nickles, T. (1980). Can Scientific Constraints be Violated Rationally? In T. Nickles (Ed.) *Scientific Discovery, Logic, and Rationality*. (285-315). Dordrecht: Reidel.

Niklasson, L., van Gelder, T. (1994). Can Connectionist Models Exhibit Non-Classical Structure Sensitivity? In A. Ram, K. Eiselt (Eds.) *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.* (664-669). Erlbaum.

Nilsson, N. J. (1991). Logic and Artificial Intelligence. *Artificial Intelligence, 47(*1-3), 31-56.

Norman, D. A. (1986). Reflections on Cognition and Parallel Distributed Processing. In J. L. McClelland, D. E. Rumelhart (Eds.) *Parallel Distributed Processing. Vol. 2: Psychological and Biological Models.* (531-546). MIT.

Norman, D. A. (1991). Approaches to the Study of Intelligence. *Artificial Intelligence, 47*(1-3), 327-346.

Norman, D. A., Draper, S. W. (1986). *User Centered System Design*: *New Perspectives on Human Computer Interaction* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Nutter, J. T. (1991). Focus of Attention, Context, and the Frame Problem. In K. M. Ford, P. J. Hayes (Eds.) *Reasoning Agents in a Dynamic World: The Frame Problem.* (219-230). Greenwich, CT: JAI Press.

O'Conner, T. (1994). Emergent Properties. *American Philosophical Quarterly, 31*(2), 91-104.

Olson, K. R. (1987). *An Essay on Facts.* CSLI.

Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Lloyd (Eds.) *Cognition and categorization*. (259-303). Hillsdale, NJ: Erlbaum.

Patel, M. J., Schnepf, U. (1992). Concept Formation as Emergent Phenomena. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (11-20). MIT.

Pellionisz, A. J. (1991). Geometry of Massively Parallel Neural Interconnectedness in Wired Networks and Wireless Volume Transmission. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (557-568) New York: Raven.

Pfeifer, R., Verschure, P. (1992). Distributed Adaptive Control: A Paradigm for Designing Autonomous Agents. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (21-30). MIT.

Piaget, J. (1954). *The Construction of Reality in the Child*. New York: Basic.

Piaget, J. (1962). *Play, Dreams, and Imitation in Childhood.* New York: Norton.

Piaget, J. (1970a). *Genetic epistemology.* New York: Columbia.

Piaget, J. (1970b). Piaget's Theory. In P. H. Mussen (Ed.), *Carmichael's Manual of Child Psychology*. (703-732). New York: Wiley.

Piaget, J. (1971). *Biology and Knowledge*. Chicago: University of Chicago Press.

Piaget, J. (1977). The Role of Action in the Development of Thinking. In W. F. Overton, J. M. Gallagher (Eds.) *Knowledge and Development: Vol. 1.* (17-42). New York: Plenum.

Piaget, J. (1985). *The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development*. Chicago: University of Chicago Press. Translated: T. Brown and K. Thampy. (Originally published 1975).

Piaget, J. (1987). *Possibility and Necessity*. *Vols. 1 and 2*. Minneapolis: U. of Minnesota Press.

Piattelli-Palmarini, M. (1980). *Language and Learning.* Cambridge: Harvard University Press.

Pierce, D., Kuipers, B. (1990). Learning Hill-Climbing Functions as a Strategy for Generating Behaviors in a Mobile Robot. In *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*. (327-336). MIT.

Pinker, S. (1988). On Language and Connectionism: Analysis of a parallel distributed processing model of language acquisition. In S. Pinker, J. Mehler (Eds.) *Connections and Symbols.* (73-193). Cambridge: MIT.

Pinker, S., Mehler, J. (1988). *Connections and Symbols.* Cambridge: MIT.

Pollack, J. B. (1990). Recursive Distributed Representations. *Artificial Intelligence, 46*, 77-105.

Pollack, J. B. (1991). The Induction of Dynamical Recognizers. *Machine Learning, 7*, 227-252.

Popkin, R. H. (1979). *The History of Scepticism.* Berkeley: University of California Press.

Popper, K. (1965). *Conjectures and Refutations*. New York: Harper & Row.

Popper, K. (1972). *Objective Knowledge*. London: Oxford Press.

Port, R., van Gelder, T. J. (in press). *Mind as Motion: Dynamics, Behavior, and Cognition*. MIT.

Posner, M. I. (1989). *Foundations of Cognitive Science.* MIT.

Priest, G. (1987). *In Contradiction*. Kluwer Academic.

Putnam, H. (1975). The meaning of meaning. In K. Gunderson (Ed.) *Language, mind, and knowledge*. (131-193). Minneapolis: University of Minnesota Press.

Putnam, H. (1990). *Realism with a Human Face*. Cambridge, MA: Harvard University Press.

Putnam, H. (1992). *Renewing Philosophy*. Cambridge, MA: Harvard University Press.

Pylyshyn, Z. (1984). *Computation and Cognition.* MIT.

Pylyshyn, Z. (1987). *The Robot's Dilemma*. Ablex.

Quartz, S. R. (1993). Neural Networks, Nativism, and the Plausibility of Constructivism. *Cognition, 48*(3), 223-242.

Quillian, M. R. (1968). Semantic Memory. In Minsky, M. (Ed.) *Semantic Information Processing* (227-270). Cambridge: MIT Press.

Quine, W. V. O. (1966a). Implicit Definition Sustained. In W. V. O. Quine (Ed.) *The Ways of Paradox*. (195-198). New York: Random House.

Quine, W. V. O. (1966b). Variables explained away. In W. V. O. Quine (Ed.) *Selected logic papers*. (227-235). New York: Random House.

Rescher, N. (1980). *Scepticism*. Totowa, NJ: Rowman and Littlefield.

Resnik, M. (1981). Mathematics as a science of patterns: Ontology and reference. *Nous*, *15*(4), 529-550.

Rich, E., Knight, K. (1991). *Artificial Intelligence*. New York: McGraw-Hill.

Richard, M. (1983). Direct Reference and ascriptions of belief. *Journal of Philosophical Logic, 12*(4), 425-452.

Ricoeur, P. (1977). The Model of the Text: Meaningful Action Considered as a Text. In F. Dallmayr & T. McCarthy (Eds.) *Understanding and Social Inquiry*. (316-334) Notre Dame: U. of Notre Dame Press.

Rivest, R. L., Schapire, R. E. (1987). Diversity-based inference of finite automata. *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, (78-87). Los Angeles, Oct. 1987.

Rivest, R. L., Schapire, R. E. (1989). Inference of finite automata using homing sequences. *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, Seattle, May 1989.

Roberts, A., Bush, B. M. H. (1981). *Neurones without Impulses*. Cambridge University Press.

Rogers, H. (1967). *Theory of recursive functions and effective computability.* New York: McGraw Hill.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton.

Rorty, R. (1982). *Consequences of Pragmatism*. Minnesota.

Rorty, R. (1987). Pragmatism and Philosophy. In K. Baynes, J. Bohman, T. McCarthy (Eds.) *After Philosophy: End or Transformation*? (26-66). MIT.

Rorty, R. (1989). *Contingency, Irony, and Solidarity*. Cambridge.

Rorty, R. (1991a). *Objectivity, Relativism, and Truth*. *Philosophical Papers vol. 1*. Cambridge.

Rorty, R. (1991b). *Essays on Heidegger and Others*. *Philosophical Papers vol. 2*. Cambridge.

Rosenbloom, P. S., Laird, J. E., Newell, A., McCarl, R. (1991). A Preliminary Analysis of the SOAR Architecture as a Basis for General Intelligence. *Artificial Intelligence, 47*(1-3), 289-325.

Rosenbloom, P. S., Newell, A., Laird, J. E. (1991). Toward the Knowledge Level in SOAR: The Role of the Architecture in the Use of Knowledge. In K. VanLehn (Ed.) *Architectures for Intelligence*. (76-111) Erlbaum.

Rosenbloom, P., Laird, J., Newell, A. (1988). Meta-Levels in SOAR. In P. Maes, D. Nardi (Eds.) *Meta-Level Architectures and Reflection.* (227-240). Elsevier.

Rosenfeld, A. (1968). *Algebraic Structures*. San Francisco: Holden-Day.

Rosenschein, S. (1985). Formal theories of knowledge in AI and robotics. *New Generation Computing, 3*, 345-357.

Rosenschein, S. J., Kaelbling, L. P. (1986). The synthesis of digital machines with provable epistemic properties. In J. Y. Halpern (Ed.) *Reasoning About Knowledge*. (83-98). Los Altos, California: Morgan Kaufmann.

Rosenthal, S. B. (1983). Meaning as Habit: Some Systematic Implications of Peirce's Pragmatism. In E. Freeman (Ed.) *The Relevance of Charles Peirce*. (312-327). Monist.

Rosenthal, S. B. (1987). Classical American Pragmatism: Key Themes and Phenomenological Dimensions. In R. S. Corrington, C. Hausman, T. M. Seebohm (Eds.) *Pragmatism Considers Phenomenology*. (37-57). Washington, D.C.: University Press.

Rosenthal, S. B. (1990). *Speculative Pragmatism.* Open Court.

Rosenthal, S. B. (1992). Pragmatism and the Reconstruction of Metaphysics: Toward a New Understanding of Foundations. In T. Rockmore, B. J. Singer *Antifoundationalism Old and New*. (165-188) Temple University Press.

Rosenthal, S. B., Bourgeois, P. L. (1980). *Pragmatism and Phenomenology: A Philosophic Encounter*. Grüner.

Rumelhart, D. E. (1989). The Architecture of Mind: A Connectionist Approach. In M. I. Posner (Ed.) *Foundations of Cognitive Science.* (133-160). MIT.

Rumelhart, D. E., McClelland, J. L. (1986). *Parallel Distributed Processing. Vol. 1: Foundations*. MIT.

Rumelhart, D. E., Norman, D. (1985). Representation of Knowledge. In A. M. Aitkenhead, J. M. Slack (Eds.) *Issue in Cognitive Modeling.* (15-62). Erlbaum.

Russell, B. (1985). *The Philosophy of Logical Atomism.* Open Court.

Ryle, G. (1949). *The Concept of Mind*. New York: Barnes & Noble.

Sacerdoti, E. (1977). *A Structure for Plans and Behavior*. Amsterdam. North Holland.

Schank, R. C., Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Erlbaum.

Scharrer, B. (1987). Evolution of Intercellular Communication Channels. In B. Scharrer, H.-W. Korf, H.-G. Hartwig (Eds.) *Functional Morphology of Neuroendocrine Systems.* (1-8). Berlin: Springer-Verlag.

Schön, D. (1983). *The Reflective Practitioner*. New York: Basic Books.

Searle, J. R. (1969). *Speech Acts***.** London: Cambridge University Press.

Searle, J. R. (1981). Minds, Brains, and Programs. In J. Haugeland (Ed.) *Mind design*. (282-306). Cambridge: MIT.

Searle, J. R. (1990). Consciousness, Explanatory Inversion, and Cognitive Science. *Behavioral and Brain Sciences, 13*, 585-642.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge: MIT.

Sejnowski, T. J. (1986). Open Questions about Computation in Cerebral Cortex. In J. L. McClelland, D. E. Rumelhart (Eds.) *Parallel Distributed Processing. Vol. 2: Psychological and Biological Models.* (372-389). MIT.

Shanon, B. (1987). On the Place of Representations in Cognition. In Perkins, D. N., Lockhead, J., Bishop, J. C. (Eds.) *Thinking: The Second International Conference*. (33-49). Hillsdale, NJ: Erlbaum.

Shanon, B. (1988). Semantic Representation of Meaning. *Psychological Bulletin, 104*(1), 70-83.

Shanon, B. (1992). Are Connectionist Models Cognitive? *Philosophical Psychology, 5*(3), 235-255.

Shanon, B. (1993). *The Representational and the Presentational.* Hertfordshire, England: Harvester Wheatsheaf.

Shastri, L., & Ajjanagadde, V. (1993). From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables, and Dynamic Bindings Using Temporal Synchrony. *Behavioral and Brain Sciences, 16*, 417-494.

Sheard, M. (1994). A Guide to Truth Predicates in the Modern Era. *Journal of Symbolic Logic, 59*(3), 1032-1054.

Shepard, G. M. (1981). Introduction: The Nerve Impulse and the Nature of Nervous Function. In A. Roberts & B. M. H. Bush (Eds.) *Neurones without Impulses.* (1-27). Cambridge University Press.

Siegelbaum, S. A., Tsien, R. W. (1985). Modulation of Gated Ion Channels as a Mode of Transmitter Action. In D. Bousfield (Ed.) *Neurotransmitters in Action.* (81-93) Amsterdam: Elsevier.

Simmons, R. F. (1973). Semantic Networks: Their Computation and Use for Understanding English Sentences. In Schank, R. C. & Colby, K. M. (Ed.) *Computer Models of Thought and Language* (69-113). San Francisco: Freeman.

Slater, B. H. (1988). Hilbertian Reference. *Nous, 22*, 283-297.

Slezak, P. (1992). Situated Cognition: Minds in Machines or Friendly Photocopiers? *Proceedings of "The Science of Cognition"* Santa Fe, New Mexico.

Slezak, P. (1994). Situated Cognition: Empirical Issue, 'Paradigm Shift', or Conceptual Confusion? In A. Ram, K. Eiselt (Eds.) *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. (806-811). Erlbaum.

Smith, B. (1986). The Link from Symbols to Knowledge. In Z. W. Pylyshyn, W. Demopoulos (Eds.) *Meaning and Cognitive Structure* (40-50). Ablex.

Smith, B. C. (1985). Prologue to "Reflections and Semantics in a Procedural Language" In R. J. Brachman, H. J. Levesque (Eds.) *Readings in Knowledge Representation*. (31-40). Los Altos, CA: Morgan Kaufmann.

Smith, B. C. (1987). *The Correspondence Continuum*. Stanford, CA: Center for the Study of Language and Information, CSLI-87-71.

Smith, B. C. (1988). The Semantics of Clocks. In J. H. Fetzer (Ed.) *Aspects of Artificial Intelligence*. (3-31). Kluwer Academic.

Smith, B. C. (1991). The Owl and the Electric Encyclopedia. *Artificial Intelligence*, 47(1-3), 251-288.

Smith, J. E. (1987). The Reconception of Experience in Peirce, James, and Dewey. In R. S. Corrington, C. Hausman, T. M. Seebohm (Eds.) *Pragmatism Considers Phenomenology.* (73-91). Washington, D.C.: University Press.

Smithers, T. (1992). Taking Eliminative Materialism Seriously: A Methodology for Autonomous Systems Research. In Francisco J. Varela & Paul Bourgine (Eds.) *Toward a Practice of Autonomous Systems*. (31-40). MIT Press.

Smithers, T. (1994). On Behaviour as Dissipative Structures in Agent-Environment System Interaction Spaces. Presented at the meeting "Prerational Intelligence: Phenomenology of Complexity in Systems of Simple Interacting Agents," November 22-26, 1993, part of the Research Group, "Prerational Intelligence," Zentrum für Interdisziplinäre Forschung (ZiF), University of Bielefeld, Germany, 1993/94.

Smolensky, P. (1986). Information Processing in Dynamical Systems: Foundations of Harmony Theory. In Rumelhart, D. E., McClelland, J. L. (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. (194-281). Cambridge, MA: MIT.

Smolensky, P. (1988). On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Smolensky, P. (1990). Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems. *Artificial Intelligence, 46*, 159-216.

Steels, L. (1991). Towards a Theory of Emergent Functionality. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (451-461). MIT.

Steels, L. (1994). The Artificial Life Roots of Artificial Intelligence. *Artificial Life, 1*(1), 75-110.

Stefik, M. (1986). The Next Knowledge Medium. *AI Magazine, 11*(3), 34-46.

Steier, D. D., Laird, J. E., Newell, A., Rosenbloom, P. S., Flynn, R. A., Golding, A., Polk, T. A., Shivers, O. G., Unruh, A., Yost, G. R. (1987). Varieties of Learning in SOAR. *Proceedings of the Fourth International Workshop on Machine Learning*. (300-311) Los Altos, CA: Morgan Kaufman.

Stein, L. (1991). An Atemporal Frame Problem. In K. M. Ford, P. J. Hayes (Eds.) *Reasoning Agents in a Dynamic World: The Frame Problem*. (219-230). Greenwich, CT: JAI Press.

Stewart, J. (1992). LIFE = COGNITION: The Epistemological and Ontological Significance of Artificial Life. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (475-483). MIT.

Stroud, B. (1984). *The Significance of Philosophical Scepticism*. London: Oxford University Press.

Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human Machine Communication.* Cambridge: Cambridge University Press.

Suppe, F. (1977a). The Search for Philosophic Understanding of Scientific Theories. In F. Suppe (Ed.) *The Structure of Scientific Theories.* (3-241). University of Illinois Press.

Suppe, F. (1977b). Afterward — 1977. In F. Suppe (Ed.) *The Structure of Scientific Theories*. (617-730). University of Illinois Press.

Sussman, G. J. (1975). *A Computer Model of Skill Acquisition.* New York: Elsevier.

Tarski, A. (1956). *Logic, Semantics, Metamathematics.* London: Oxford.

Tate, A. (1974). INTERPLAN: A Plan Generation System that can deal with Interactions between Goals. Memorandum MIP-R-109. Machine Intelligence Research Unit. Edinburgh. University of Edinburgh.

Tennant, H. (1980). Evaluation of Natural Language Processors. Technical Report T-103 and Ph.D. Thesis. Coordinated Science Laboratory, University Of Illinois.

Terveen, L. G. (1993). Intelligent Systems as Cooperative Systems. *International Journal of Intelligent Systems* (*Special Issue on The Social Context of Intelligent Systems*). *3*(2-4), 217-249.

Tesar, B. B., Smolensky, P. (1994). Synchronous Firing Variable Binding is a Tensor Product Representation With Temporal Role Vectors. In A. Ram, K. Eiselt (Eds.) *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.* (870-875). Erlbaum.

Thatcher, R. W., John, E. R. (1977). *Functional Neuroscience Vol. 1 Foundations of Cognitive Processes.* Hillsdale, NJ: Erlbaum.

Thayer, H. S. (1973). *Meaning and Action*. Bobbs-Merrill.

Toth, J. A. (in press). Review of *Reasoning Agents in a Dynamic World: The Frame Problem. Artificial Intelligence*.

Touretzky, D. S., Pomerleau, D. A. (1994). Reconstructing Physical Symbol Systems. *Cognitive Science, 18*(2), 345-353.

Turner, C. D., Bagnara, J. T. (1976). *General Endocrinology*. Philadelphia: Saunders.

Ullman, S. (1980). Against direct perception. *The Behavioral and Brain Sciences, 3,* 373-381.

Valiant, L. (1984). A Theory of the Learnable. *Communications of the ACM*. 27: 1134-1142.

van Gelder, T. J. (1990). Compositionality: A Connectionist Variation on a Classical Theme. *Cognitive Science, 14*(3), 355-384.

van Gelder, T. J. (1991). What is the "D" in "PDP"? A Survey of the Concept of Distribution. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.), *Philosophy and Connectionist Theory.* (33-59). Erlbaum.

van Gelder, T. J. (1992). Tutorial: New Directions in Connectionism. *Society for Philosophy and Psychology*, McGill University.

van Gelder, T. J. (in press-a). What Might Cognition be if not Computation? In R. Port, T. J. van Gelder (Eds.) *Mind as Motion: Dynamics, Behavior, and Cognition*. MIT.

van Gelder, T. J. (in press-b). Defining "Distributed Representation." *Connection Science*.

van Gelder, T. J., Port, R. (1994). Beyond Symbolic: Towards a Kama-Sutra of Compositionality. In V. Honavar & L. Uhr (Eds.) *Symbol Processing and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration*. (107-125). San Diego: Academic Press.

van Gelder, T. J., Port, R. (in press). It's About Time: An Overview of the Dynamical Approach to Cognition. In R. Port, T. J. van Gelder (Eds.) *Mind as Motion: Dynamics, Behavior, and Cognition.* MIT.

van Gelder, T., Niklasson, L. (1994). Classicalism and Cognitive Architecture. In A. Ram, K. Eiselt (Eds.) *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.* (905-909). Erlbaum.

Van Gulick, R. (1982). Mental Representation: A Functionalist View. *Pacific Philosophical Quarterly,* 3-20.

Vera, A. H., Simon, H. A. (1993). Situated action: A symbolic interpretation. *Cognitive Science, 17*(1), 7-48.

Vera, A. H., Simon, H. A. (1994). Reply to Touretzky and Pomerleau: Reconstructing Physical Symbol Systems. *Cognitive Science, 18*(2), 355-360.

Visser, A. (1989). Semantics and the Liar Paradox. In D. Gabbay, F. Guenther (Eds.) *Handbook of Philosophical Logic.* (617-706). Dordrecth: Reidel.

Vizi, E. S. (1984). *Non-synaptic Transmission Between Neurons: Modulation of Neurochemical Transmission.* Wiley.

Vizi, E. S. (1991). Nonsynaptic Inhibitory Signal Transmission Between Axon Terminals: Physiological and Pharmacological Evidence. In K. Fuxe & L. F. Agnati (Eds.) *Volume Transmission in the Brain: Novel Mechanisms for Neural Transmission*. (89-96) New York: Raven.

von Glasersfeld, E. (1979). Radical constructivism and Piaget's concept of knowledge. In F. B. Murray (Ed.) *The impact of Piagetian theory.* (109-122). Baltimore: University Park Press.

von Glasersfeld, E. (1981). The concepts of adaptation and viability in a radical constructivist theory of knowledge. In I. E. Sigel, D. M. Brodzinsky, R. M. Golinkoff (Eds.) *New Directions in Piagetian Theory and Practice*. (87-95). Hillsdale, NJ: Erlbaum.

Vygotsky, L. S. (1962). *Thought and Language.* Cambridge: MIT Press.

Waltz, D. (1982). The State of the Art in Natural-Language Understanding. In Lehnert, W., & Ringle, M. (Ed.) *Strategies for Natural Language Processing.* (3-32). Hillsdale, NJ: Erlbaum.

Waltz, D., Feldman, J. A. (1988a). Connectionist Models and Their Implications. In D. Waltz, J. A. Feldman (Eds.) *Connectionist Models and Their Implications*. (1-12). Norwood, NJ: Ablex.

Waltz, D., Feldman, J. A. (1988b). *Connectionist Models and Their Implications*. Norwood, NJ: Ablex.

Ward, R. S., Wells, R. O. (1990). *Twistor Geometry and Field Theory.* Cambridge.

Warner, F. (1983). *Foundations of Differentiable Manifolds and Lie Groups.* Springer-Verlag.

Warnke, G. (1987). *Gadamer: Hermeneutics, Tradition, and Reason*. Stanford University Press.

Warren, D. H. (1974). WARPLAN: A System for Generating Plans. DCL Memo 76. Department of Artificial Intelligence. Edinburgh. University of Edinburgh.

Wertsch, J. L. (1985). *Culture, Communication, and Cognition: Vygotskian Perspective*s. Cambridge: Cambridge University Press.

Whorf, B. L. (1956). *Thought, Language, and Reality*. Cambridge: MIT Press.

Wiggins, S. (1990). *Introduction to Applied Nonlinear Dynamical Systems and Chaos.* Springer-Verlag.

Wilkins, D. E. (1988). *Practical Planning: Extending the Classical AI Planning Paradigm*. San Mateo, CA: Morgan Kauffmann.

Wilks, Y. (1982). Some Thoughts on Procedural Semantics. In Lehnert, W. & Ringle, M. (Ed.) *Strategies for Natural Language Processing*. (495-516). Hillsdale, NJ: Erlbaum

Wimsatt, W. C. (1972). Teleology and the Logical Structure of Function Statements. *Studies in the History and Philosophy of Science, 3*, 1-80.

Winograd, T. (1972). *Understanding Natural Language***.** New York. Academic Press.

Winograd, T. (1975). Frame representations and the declarative-procedural controversy. In D. G. Bobrow & A. Collins (Eds.) *Representation and understanding*. (185-210). New York: Academic Press.

Winograd, T. (1976). Towards a Procedural Understanding of Semantics. *Revue Internationale du Philosophie 3*(3-4).

Winograd, T., Flores, F. (1986). *Understanding Computers and Cognition.* Norwood, NJ: Ablex.

Wittgenstein, L. (1961). *Tractatus Logico-Philosophicus*. New York: Routledge.

Woods, W. A. (1986). Problems in Procedural Semantics. In Z. Pylyshyn, W. Demopoulos (Eds.) *Meaning and Cognitive Structure: Issues in the Computational Theory of Mind*. (55-85). New York: Springer-Verlag.

Woods, W. A. (1987). Don't Blame the Tools. *Computational Intelligence, 3*, 228-237.

Wright, L. (1973). Functions. *Philosophical Review, 82*, 139-168.

Yamauchi, B. M., Beer, R. D. (1994). Sequential Behavior and Learning in Evolved Dynamical Neural Networks. *Adaptive Behavior, 2*(3), 219-246.

Yaqub, A. M. (1993). *The Liar Speaks the Truth.* Oxford.

Index

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z