AUTONOMY, FUNCTION, AND REPRESENTATION

Mark H. Bickhard
Cognitive Science, Lehigh University
Bethlehem, PA 18015

Abstract

Autonomy is modeled in terms of the property of certain far-from-equilibrium open systems to contribute toward maintaining themselves in their far-from-equilibrium conditions. Such contributions in *self-maintenant* systems, in turn, constitute the emergence of nonnative *function*. The intrinsic thermodynamic asymmetry between equilibrium and far-from-equilibrium processes yields the intrinsic normative asymmetry between function and dysfunction. Standard etiological models of function render function as causally epiphenomenal, while this model is of the emergence of causally efficacious function. *Recursive* self-maintenance — the meta-property of maintaining the property of being self-maintenant across variations in environment — yields the emergence of *representation*. This model of representation satisfies multiple criteria that standard approaches — such as symbolic or connectionist, or those of Fodor, Dretske, or Millikan — cannot.

* * *

Naturalism is the regulatory assumption that the world is an integrated reality, that there are no locally specific dedicated substances or properties or processes involved in, or required for understanding, the world. Naturalism is a fundamental rejection of ad-hoc ontologies, postulates, and explanations — of levels of explanation beyond which no further questions can be asked.

Accounting for the nature and emergence of representation is one of the primary challenges still facing naturalism. Other phenomena — such as fire, magnetism, and life — we understand, at least in principle, as natural phenomena, and no longer feel compelled to postulate ad-hoc substances or fluids to explain them. But mind and mental phenomena, with representation central among them, still elude a naturalistic account.

I will present a model of the natural emergence of representation in certain kinds of far-from-equilibrium systems. In particular, some far-from-equilibrium systems manifest an *autonomy* with respect to their environments, and the model will exhibit the emergence of normative *function* in a relatively simple form of autonomy, and of *representation* in a stronger form of autonomy.

Emergence

First, a word about emergence itself. There are a number of approaches to modeling representation (and function) that construe the issue to be one of how we talk about things: explicating the conditions under which it would be appropriate to talk about representation is all there is to the matter (e.g., Clark, 1997). Such an avoidance of ontological concerns yields a non-naturalistic outcome: representation in a system is modelable only in terms of the attributions, the glosses, of some other epistemic agent — that is, some other agent that itself makes use of representation, or could be glossed as doing so. This is either circular or it initiates a regress of glossing representations in agents that have glossed representations in still other agents, and so on. If representation is to be a natural part of the world, it needs to be modeled in a way that makes sense of its ontological emergence, not just as a form of gloss.

Furthermore, that emergence must be of a causally efficacious property. A technically emergent representation that was nevertheless causally epiphenomenal would not offer an account of how representation makes a difference in the processes of the world. It would be non-natural in the sense of being true but superfluous, like the fact that three stars happen to be co-linear from the perspective of the current position of the earth (or the constellations in general).

There is a major challenge to any claim of non-epiphenomenal emergence that I would like to mention and address briefly. Kim (1993, 1997, 1998) has argued that we are faced with a dilemma: either 1) *all* of causality is resident at the level of the most fundamental particles, and any higher level regularities are simply the working out of the particle interactions with the relevant initial and boundary conditions, in which case any higher level regularities (including those of representation and mind in general) are causally epiphenomenal, causally superfluous, relative to the particle level, or 2) the world is not physically closed, that is, interactions at the most basic physical level do not suffice to determine what happens in the world, and something else at higher levels is required, something that sounds a lot like a dualism, or worse. This dilemma is not a happy one. It pits naturalism against emergence, and it appears that one or the other has to lose.

I find nothing to criticize in the steps of this argument. It is valid. But, I have argued that it is unsound; it is based on a false premise (Bickhard, 1998; Bickhard & Campbell, 2000). In particular, it is based on a particle metaphysics, and our best contemporary science tells us that there are no particles, only processes. If so, then we need a process metaphysics, and the argument does not go through on the premise of a process metaphysics.

First, to the point that there are no particles. Quantum field theory shifts the basic ontology of the universe from particles to quantum fields (Aitchison, 1985; Aitchison & Hey, 1989; Brown & Harré, 1988; Davies, 1984; Ryder, 1985;

Sciama, 1991; Weinberg, 1977, 1995). Particle-like processes and interactions are the result of the quantization of field processes and interactions, and those are no more particles than are the integer number of oscillatory waves in a guitar string. Everything is quantized field processes.

Second, to the point that Kim's argument does not go through in a process framework. The key difference between particles and processes in this regard is that ultimate particles have no structure or organization of their own. They may *participate* in interactions with other particles in some overall organization, but a single particle is point-like, and cannot have any internal structure. That makes any properties or regularities of the organization "merely" the stage upon which and within which the particles work out their interactions. In particular, there is no justification for modeling the organization itself as having any causal power: that is inherent in the particles.

Processes, in contrast, exist only in some organization or another. There is no such thing as process with no organization. There is no level beyond which or below which organization is left behind. The notion of a point process is incoherent. Anything that has causal power, therefore, will have causal power as an organized feature of the world. Furthermore, in general, different organizations yield different causal properties, so organized process is a legitimate, even necessary, locus of causal power — unlike for the case of a particle metaphysics. Still further, there is no level above which we can ignore the possibility of new causally efficacious properties inhering in new organization. For one counterexample to any such assumption, quantum effects can occur at any scale, e.g., superconductivity. So we cannot simply assign causal power to organization below some special scale.

All scales of organization of process are candidates for the non-epiphenomenal emergence of new causally efficacious power. Some of those scales, and some of those organizations, perhaps, may model the emergent properties of representation and other mental phenomena.

Forms of Stability

Some organizations of processes are fleeting, such as the fall of a nut from a tree. Others are stable, and may persist in the same organization for eons, such as an atom (at least under most terrestrial conditions). Stability over time and against perturbation manifest a *cohesion* of an overall organization of process (Christensen, Collier, Hooker, 1999; Christensen & Hooker, 1998, 2001; Collier, 1988, 1999). There are two fundamental forms of process stability: 1) energy well stability, and 2) far-from-equilibrium stability. Energy well stability is exhibited when some organization of process requires energy to change its organization and the ambient environment does not impinge that level of energy into or onto the process. Atoms are straightforward examples: they are a furious process of electron waves around an even more furious dance of quarks and gluons. If sufficient energy is introduced into this process, it is altered and perhaps disintegrates; it loses its cohesion.

Far-from equilibrium stability occurs in a process organization that is not at thermodynamic equilibrium and for which the persistence of that organization requires that the process not move to equilibrium. A candle flame, for example, manifests a short term stability and persistence, but only so long as fuel and oxygen are input to the process. The maintenance of far-from-equilibrium processes in their far-from-equilibrium conditions requires transactions with the environment, otherwise they would move toward equilibrium and the organization would cease. That is, far-from-equilibrium processes that exhibit stability are necessarily open processes.

Autonomy

Cohesion, in a very general sense, is the property of stability against perturbations. Energy well process organizations, such as an atom or a rock, exhibit a fundamental form of cohesion. But more interesting and more important for current purposes are the kinds of cohesion that some far-from-equilibrium systems can exhibit, in which they can make active contributions to their own stability; they exhibit *autonomy* in the sense of actively contributing to their own persistence (Christensen, Collier, Hooker, 1999; Christensen & Hooker, 1998, 2001, in press; Collier, 1999). Autonomy in this sense is a graded concept: there are differing kinds and degrees of such "active contributions." I will address first what I have called self-maintenant systems.

Some far-from-equilibrium processes are completely dependent for their continued existence on continued external sources of support. A chemical bath, for example, in which perhaps interesting far-from-equilibrium processes are taking place, is dependent on the external pumps continuing to pump the necessary solutions into the bath, and on the external reservoirs of those solutions continuing to contain sufficient quantities of those solutions. Such a far-from-equilibrium process exhibits at best a minimal autonomy.

A candle flame, in contrast, makes several active contributions to its own persistence. It maintains above combustion threshold temperature. It vaporizes wax into a continuing supply of fuel. In a standard atmosphere and gravitational field, it induces convection, which pulls in continuing oxygen and removes combustion products. A candle flame, in other words, tends to maintain itself; it exhibits *self-maintenance* (Bickhard, 1993). Self-maintenance abilities have limitations — the candle flame cannot seek more fuel when the candle is almost gone, for example — but they do contribute to the maintenance of the conditions for their own existence; they succeed in countering the impact or development of conditions and processes that could otherwise destroy their far-from-equilibrium persistence, such as the local depletion of oxygen for the candle flame.

Some far-from-equilibrium processes, however, can do better than that.

They can not only deploy processes that tend to contribute to self-maintenance — the generation of heat by the flame — but they can also deploy *different* processes depending on differences in the environmental circumstances that they face. Candle flames cannot do this, but if they could search for fuel when the candle was low, that would be an example. A non-science fiction example would be a bacterium that can swim up a sugar gradient, but tumble if it happens to be swimming down a sugar gradient (Campbell, D. T., 1974, 1990). Swimming and tumbling are two different interactions that are appropriate in the sense of contributing to self-maintenance in differing conditions, and the bacterium can switch between them appropriately (usually) as the conditions change.

Such systems exhibit a kind of maintenance of their own abilities to be self-maintenant. They shift their self-maintenant processes so as to maintain self-maintenance as the environment shifts. They exhibit a *recursive self-maintenance* (Bickhard, 1993), in which their interactions with their environments exhibit a causal closure with the maintenance of the conditions in the system for those very interactions (Christensen, Collier, Hooker, 1999; Christensen & Hooker, 1998, 2001, in press; Collier, 1999). This is a much stronger form of autonomy in that stability is maintainable not only in certain ranges of conditions, but also within certain ranges of *changes* in conditions.

Recursive self maintenance requires some sort of infrastructure in the system that engages in the relevant shifts of system processes — some sort of switching mechanism. Infrastructure, in this sense, is structure in the system that is stable relative to the time scales in which the switching takes place: the internal cellular structure in the bacterium, for example. Infrastructure could be stable in an energy-well sense, but more commonly will also be far from equilibrium, but with a longer time scale process of replacement and recreation. That is, with a longer time scale dedicated metabolism (Moreno & Ruiz-Mirazo, 1999).

This infrastructure will exhibit both energetic and informational aspects. The informational aspects have to do with the accomplishment of the process switching, and will be addressed in greater detail below. The energetic aspects are concerned both with the accomplishment of the switching and with the accomplishment of the processes to which the system can be switched. For example, the bacterium expends appreciable energy in both swimming and tumbling, though it is directed in importantly different ways in the two cases.

The energy directing aspect of infrastructure suggests a kind of autonomy that is in between self maintenance and recursive self maintenance. There is not necessarily any infrastructure in a simply self maintenant system, such as a candle flame. The infrastructure in a recursively self maintenant system has both energetic and informational aspects. In between would be a system with infrastructure that contributed to self maintenance via the directing of energy, the accomplishment of relevant work, but with no alternatives, no switching, and, therefore, no relevant informational aspects. An example might be primitive cells that, say, metabolize sulfur and have appropriate infrastructure for doing so, but which have no alternatives available, and thus no need to switch among alternatives. Self maintenant systems are autonomous in the sense that they contribute to their own persistence. Systems with energy directing infrastructure are autonomous in the stronger sense that they direct work toward their own persistence. This constitutes the basic minimal form of autonomy as modeled by Christensen, Hooker, and Collier (Christensen, Collier, Hooker, 1999; Christensen & Hooker, 1998, 2001, in press; Collier, 1999). Recursive self maintenant systems exhibit a still stronger kind of autonomy in which the infrastructure engages in process switching as well as energy directing. I will argue that this grading of autonomy is a grading that is relevant to function and representation.[1]

Self-Maintenance and Function

The contributions that a self-maintenant system make to its own continued existence are, in that sense, *functional* for that system: They serve the general function of helping to maintain the existence of the far-from-equilibrium system. Derivatively, any components or parts of a system, perhaps the tumbling machinery for a bacterium, serve such a function insofar as they make such a contribution. Serving a function, in this sense, is necessarily relative to the system whose maintenance is being contributed to. The gut of a parasite will serve functions for the parasite, but be dysfunctional for the host. In this model, the intrinsic thermodynamic asymmetry between far-from-equilibrium and equilibrium processes yields the intrinsic *normative asymmetry* between function and dysfunction.

A part or aspect of a system will *have a function* insofar as it is an instance of a type which tends to serve, has a disposition to serve, that function for the type of the overall system. Having a function, then, depends on what types the system and subsystem belong to. In practice, such typification generally depends on a part having the relevant infrastructure to belong to a type, even if the part doesn't in fact serve the function at issue — even if it is in that sense dysfunctional. A kidney, then, may have the function of, say, filtering blood, even if this particular kidney doesn't in fact filter blood.

Such issues of the assignment of types are important, but the ontology of having a function cannot depend ultimately on assignment, or gloss, on pain of the same sort of circularities and regresses as mentioned in the discussion of emergence. This issue concerning having a function, in fact, is "just" a special case of the same general point about emergence. The sense in which a mass of tissue *is a kidney*, then, and therefore *has the function of filtering blood* — whether or not it actually does so — must be non-epiphenomenally emergent.

Part of the importance of having the proper infrastructure to be a kidney, and therefore to have the function of filtering blood, is how close this kidney is to being able to serve that function. Perhaps it can do so under some conditions but not others, or under special but attainable conditions, or rare conditions, or with some help — perhaps a certain nutrient — or a drug or surgery, and so on. In contrast, the special cases or counterfactuals that would be involved in making it possible for this scar tissue — located where a kidney used to be — to be able to filter blood are quite distant, and the sense in which that scar tissue has the function of filtering blood, even though it currently does not, is similarly distant.

Another causally relevant issue is that systems of the including type — complex organisms for the case of kidneys — *must* have some functions served, such as filtering blood, if they are to persist at all, not just to be able to persist under certain conditions. The type, then, specifies *as part of the nature of the type* that certain parts have the functions of making those necessary contributions to the maintenance of required far-from-equilibrium conditions for the general system to be stable. For complex organisms with circulatory systems, for example, something must filter blood, and being an instance of the type "complex organism with circulatory system" is being an instance of a type that contains typical parts whose functions are to filter blood. If there is nothing in fact serving that function, then the organism will cease. If there is a part of the organism that is of the type kidney, but this particular kidney does not filter blood, then this particular kidney has the function of filtering blood, but it does not do so; it is dysfunctional.

A phenomenon that integrates these issues concerning normative function and that provides a model of the emergence of functional types is that of *dynamic presupposition*. The basic notion is that some dynamic processes presuppose other processes or conditions in order for the given processes to be proceed successfully. If the dynamic presuppositions do not hold, then the process fails. If the heart does not pump blood, then the organism fails.

Dynamic presupposition is itself already a normative property; it is dependent on the presupposing processes being metaphysically capable of success and failure. The model of the emergence of *serving a function* provides the necessary normative framework for this normativity of dynamic presupposition: a process will succeed or fail in serving its function(s) in a self maintenant far from equilibrium system depending, in part, on whether or not its dynamic presuppositions hold.

Dynamic presupposition gives rise to a typification that grounds "having a function." Dynamic presupposition can be of particular phenomena occurring, conditions obtaining, and so on, at particular places and times. Insofar as the organization, perhaps the infrastructural organization, of the presupposing process determines some other component as having the correct infrastructural, spatial, and temporal properties, it thereby typifies that component as of the type that is presupposed to satisfy the presupposition. That component, in other words, *has the function* of satisfying the dynamic presuppositions, whether or not it actually does so (see Christensen & Bickhard, 2001, for a more detailed elaboration of this part of the model, including of "having a proper function").

This model of function as arising in far-from-equilibrium systems, and being constituted as contributions to the creation or support of required conditions for the maintenance of the far-from-equilibrium processes, is a genuine emergence. It is a property of certain kinds of open far-from-equilibrium systems, and it is causally efficacious. It makes a causal difference in the world whether or not this organism or this flame persists.

Etiological Approaches. This is in contrast to the dominant alternative account of normative function, the etiological approach (Godfrey-Smith, 1994; Millikan, 1984, 1993). The etiological approach models the having of a function as being constituted in having the right history, generally the right evolutionary history. Kidneys, for example, exist *because* their evolutionary predecessors did in fact filter blood, so this kidney has that function even if it is not serving it.

This is a strong model and appeals to strong intuitions, but it has, I argue, fatal problems. Millikan, one of the major proponents of the etiological approach (Millikan, 1984, 1993), points out that this approach means that, if a lion were to miraculously pop into existence, perhaps by the coming together of molecules from the air, that was molecule for molecule identical with a lion in the zoo, the heart of the science fiction lion would, nevertheless, have no function: it would have no evolutionary history, and, therefore, not the right evolutionary history. For a less far fetched manifestation of this issue, the first time some new contribution is made to the success of an organism, there is no function being served, because nothing has the function to serve it, because there is no evolutionary history.[2]

Millikan is willing to accept these consequences because of the other virtues of the approach, at root virtues of naturalism, of the etiological account of normative functions. I argue, however, that an etiological approach in fact fails to provide a naturalistic account of normative function. The reason is already manifest in the lion example: the reason that the molecular duplicate of the lion in the zoo nevertheless has no parts with functions is, according to an etiological approach, because nothing has the correct history. But, by assumption, the current states of the zoo lion and the science fiction lion are identical. Therefore, function, on this account, is not constituted by any property of current state of a system. That is, function is not emergent in any current state.

But only current state is capable of being causally efficacious. History, of any kind, can have causal consequence only via its influences on current state, and current state, in this approach, is not adequate for the emergence of function. So the etiological approach gives us a model of function that is causally epiphenomenal (Bickhard, 1993, 1998). That fails to provide a naturalistic account of function, unless function truly is just a matter of gloss, with no causal consequences in a system itself — excepts perhaps via the glossing individual.

The etiological approach, then, fails as a naturalistic model of function. It does so because it renders the current state of a system inadequate for the emergence of function. The etiological approach can make sense of, can explain, why a particular kind of (sub)system exists — kidneys, for example — or can perhaps explicate the notion of *"being designed"* for some function (Allen & Beckoff, 1995), but it cannot naturalistically explicate the emergent ontology of function itself.

Recursive Self-Maintenance and Representation

A self-maintenant system contributes to its own conditions of stability and persistence. A *recursive* self-maintenant system can shift among differing kinds of processes in order to maintain the property of being self-maintenant in varying conditions. I have outlined a model of function in terms of self-maintenance, and will now outline a model of representation in terms of recursive self-maintenance, a model called *interactivism.*

A recursive self-maintenant system must have some way of differentiating those environments in which it will engage in one process versus those in which it will engage in some other process. If the differing processes are in fact to be appropriately contributory to self-maintenance, then those environmental differentiations will be of conditions in which one process is appropriate versus those in which some other is appropriate. That is, if we assume that there are two relevant processes, Q and R, say, the environmental differentiations will be of environments in which Q is appropriate versus those in which R is appropriate.

The most general way in which such differentiations can take place is for the system to interact with its environment in ways in which two or more internal outcomes of the interaction are possible. The internal course of an interaction with the environment will depend in part on the (sub)system controlling the interaction, and in part on the environment being interacted with. Some environments will yield a final internal outcome of, say, A, while other environments engaged by that same subsystem may yield a final internal outcome of B. This differentiates A-type environments from B-type environments. If A-type environments also happen to be environments in which Q interactions are appropriate — tend to be self-maintaining — and B-type environments are appropriate for R-interactions, then there is an ability to differentiate environments in a way that can be used to appropriately shift among self-maintaining processes: if in an A-type environment, (it is appropriate to) do Q. The bacterium, for example, must somehow differentiate "swimming up sugar gradients" from "swimming down sugar gradients" in order to continue swimming in the first case and to tumble in the second.

A less powerful manner in which such differentiation can occur is if the "interaction" involves no outputs from the system, but, instead, it "simply" processes inputs that it receives in order to arrive at its internal differentiating states, perhaps A or B again. Such passive differentiations may or may not suffice to be useful for appropriate shifting, but, when they are adequate, they are also less costly.

Such passive differentiations are commonly taken to constitute representations themselves, particularly as they occur in the early sensory systems (e.g., Fodor, 1987, 1990, 1998; cf. Bickhard & Richie, 1983). Models of representation as originating in such sensory encodings, sensory impressions from the environment, have been with us for millennia (e.g., the analogy of representations as impressions into a waxed slate from Plato and Aristotle, Kemp, 1998), but have always ultimately failed (Bickhard, 1980, 1993; Bickhard & Richie, 1983; Bickhard & Terveen, 1995; Campbell & Bickhard, 1986). I will argue below that this approach to modeling representation cannot work. The interactive model of representation and representational content focuses instead on the process of switching among, or indicating the appropriateness of, alternative self-maintaining processes. That is, standard models construe representation as backward looking, toward what a differentiation is a differentiation of, while the interactive model construes representation as future oriented, toward the uses a system can make of those differentiations — uses toward its own self-maintenance or autonomy.

Interactive Content. The interactive model posits differentiating processes, whether fully interactive or passive input processing, as being necessary for the function of indicating which further interactions might be possible or appropriate. Standard approaches, especially information semantic approaches, construe such differentiations as already constituting representation, with the purported representational content being what has been, or "usually" is, differentiated. These approaches don't work. So, I turn to the interactive model of content.

Differentiating an A-type environment may indicate that a Q-type interaction — and, perhaps, also a T-type and an X-type — is appropriate in this just-differentiated environment. But that indication may be false. The environment may not cooperate; Q may fail. If Q fails, then the indication was false. The indication, in turn, has the form of an implicit predication *about* that environment: "this environment (A-type environments in general) are Q-type environments". So the indication makes a predication about the environment, a predication that may be true or false, and, if false, a predication whose falsity may be detectable by the system itself.

This, according to the interactive model, constitutes the most primitive emergence of representational truth value. And what is the content of that indication? At an explicit level, the content is simply "here is a Q-type environment". Implicitly, however, there will be particular properties, or combinations of properties, that would support the success of Q should Q be undertaken in the presence of some sufficient subset of that set of supporting properties. That is, the indication that this is a Q-type environment implicitly defines the class of environments in which Q would be successful, or, equivalently for current purposes, the properties that would support Q being successful. Those implicitly defined supports for Q's success are the content of indicating the (potential) success of Q.[3]

This is a model of content that does not require an observer for its definition. The indications that implicitly define such content are definable in terms of current state — they are not epiphenomenal. The content itself is, in principle, determinable by analysis of current state — though it is strictly implicit for the system per se. Error is directly system detectable. The content emerges simply and naturally in appropriate system organizations. Interactive representation is a strong candidate as a model of the nature of representation in general.

Information Semantics and Other Alternative Models. There are three major approaches to representation in the current literature that are alternatives to the interactive model that I would like to comment on. The first is that of information semantics, particularly the version proposed by Fodor. Information semantics attempts to model the nature of *representational content* — to model the nature of the specification of what a representation is supposed to represent — in terms of the representation having some sort of informational correspondence to what it is supposed to represent. Informational correspondence has several possible forms in alternative versions of the approach. It could be a causal correspondence, such as that created by light bouncing off an object and into the eye of an observer, or a lawful correspondence, such as the lawfulness of such a light propagation and patterning given the kind of object it is, or a strictly informational correspondence, such as that smoke carries information about fire, or a conventional correspondence, such as that the word "fire" carries information about the phenomenon of fire.

The interactive differentiations discussed above, particularly the passive differentiations such as in the early visual system, would, in an informational approach, be construed as being themselves representations. The internal differentiating states, A and B, for example, are in causal or informational correspondences with whatever properties in the environment in fact underlie the fact that the interaction ends in internal final state A or that it ends in B. Those internal states do carry information about the environmental properties that support the corresponding differentiations. An observer could infer from the fact that the system has arrived at A that the system is in an environment with such and such properties — if the observer knew what properties did in fact underlie arriving at internal state A. An observer, for a real case, can infer from the continued swimming of the bacterium that the bacterium is differentiating an "up a sugar gradient" environment.[4]

Information semantics approaches take some version of being in an informational correspondence as constituting being a representation. The model I am outlining takes differentiations to be no more than differentiations, and posits no content inside the system at all merely in virtue of have made a differentiation or being in a differentiating state. Differentiations per se have no content. Informational semantics claims that they do. So, a direct conflict emerges at this point. I argue that the informational account of representational content, and, thus, of representation, cannot work.

There is a multitude of relevant arguments (Bickhard, 1993; Bickhard & Terveen, 1995); I will focus on just a sampling and on three important models of representation in the literature. One important entree into this multitude has to do with the problem of the possibility of representational error. Simply, information semantics approaches have a great deal of difficulty accounting for that possibility: If the relevant informational correspondence exists, the representation exists and it is correct, while if the relevant informational correspondence does not exist, then the representation does not exist, and, therefore, it cannot be incorrect. How is representational error possible on such an account?

The etiological approach is not a paradigm instance of informational semantics, and it has a ready answer to the question about representational error. A representation represents whatever it is its proper function to represent, and it represents falsely if that content is false of a current representational target (Millikan, 1984, 1993; Cummins, 1996). So, if a content of "cow" is attributed to a horse as target, that will be false. Unfortunately, the epiphenomenality of function in this approach visits itself on the derivative model of representation, and so there is no naturalistic model of representation available here.

Fodor (1987, 1990, 1991, 1998) does work within the informational semantics approach, and his attempts to handle the problem of representational error are ingenious and revealing. The key to Fodor's model for current purposes is the notion of asymmetric dependence. If "cow" is attributed to a horse on a dark night, that is an error, according to Fodor's account, because such false attributions are asymmetrically dependent on correct attributions. That is, horses on dark nights would not evoke the cow representation if cows did not evoke it, so the "horse on a dark night" evocations are dependent on the "cow" evocations. But that dependence is asymmetric, it is not reciprocated, in the sense that cows evoking the cow representation would do so even if there never was and never would be a "horse on a dark night" evocation. So, false evocations are dependent on true evocations, but asymmetrically so. The basic intuition here is that false evocations are in some sense parasitic on true evocations.

But Fodor's model does not work either. Consider first a counter-example: a neurotransmitter docks in a receptor molecule on the surface of a cell and triggers resulting biochemical activity inside: we have nomological informational correspondence between the neurotransmitter — say, dopamine — and the cell activities. On the other hand, a poison — say, crank — can also dock on such receptors and trigger the internal cell activity. Again, we have nomological correspondence, and this time it is asymmetrically dependent on the dopamine nomological correspondences (Bickhard, 1993; Levine & Bickhard, 1999). So, the cell activities, by this model, should represent dopamine, and their evocation in response to crank should constitute representational error. But nothing of the sort is going on. Insofar as Fodor's asymmetric dependence captures error, it captures functional error just as well as it captures representational error, and it fails to distinguish between them.

Furthermore, consider what is involved in a situation in which a false representational evocation is present. The "cow" representation is evoked by a horse on a dark night, and what makes that an instance of error, — instead of, say, an evocation of a representation that represents "cows OR horses on dark nights" — is the condition of asymmetric dependency between the types of evocations. But that asymmetric dependency is a complex matter of the truth value of particular second order dependency relations among classes of counterfactuals, counterfactuals involving the various possible combinations of presence and absence of "cow" and "horse" evocations. Again, note that this structure of counterfactuals is not definable in terms of current state of the system, though in a different way than for the etiological approach. So, again we have an epiphenomenal model of representation — even if we overlook functional counterexamples such as that of dopamine and crank.

Finally, note that some organisms, at least some of the time — e.g., human beings — are capable of detecting errors in their own representations. That is not possible on either the Fodor or the etiological approach. Organisms do not have access to their own relevant evolutionary history to be able to determine what their representations are supposed to represent, nor to the relevant relationships among classes of counterfactual nomological evocations to be able to determine what their representations are supposed to represent. In neither case can an organism have access to its own representational contents. And, furthermore, even if it did, detection of error would require comparing that content with the current target — the "cow" representation with the actual horse — to determine that they did not match, but that later step is just the problem of representation all over again. These approaches make system detectable representational error impossible, but it clearly is in fact possible, so they are refuted.

For a third possibility, consider Dretske's model (Dretske, 1988). He proposes that a representation is an internal state that has been recruited via instrumental conditioning to participate in the cause of some behavior. He unpacks this in terms of that internal state carrying information about the environmental conditions that support the success of the behavior. So, "C [internal state] is recruited as a cause of M [behavior] *because* of what it indicates about F [external conditions], the conditions on which the success of M depends. Learning of this sort is a way of shaping a structure's causal properties in accordance with its indicator properties. C is, so to speak, *selected* as a cause of M because of what it indicates about F. ... C thereby becomes a representation of F." (Dretske, 1988, pg. 101).

Note first of all that this too is a kind of etiological model, but in which the relevant etiology is of a proper learning history (though evolutionary history may be involved in what counts as "the success of M"). The function of C is to indicate F, by virtue of this learning history, and that is crucial to the possibility of error: C may be evoked in circumstances that do not include F, contrary to its representational function. As an etiological model, however, it is epiphenomenal: representation is not defmable in terms of current state.

The crucial relationship in this model, however, is that of C being recruited as a cause of M *because* of what it indicates about F. That "because" cannot be a causal relationship. The learning process has no access to what C might carry information about, what it might indicate about the environment. That is a relation between C and the environment, and the learning process internal to the organism only has functional access to the state C itself, not to the existence nor the other end of any indicating relationship that C might participate in. Therefore, it cannot be the case that C is recruited in any causal sense "because of what it indicates about F". Causally, C is recruited as a cause of M because it "indicates" *the success of M.*

Dretske, however, is not attempting a causal model. Instead, he intends the "because" in an explanatory sense. That is, a full explanation of why C is recruited as a cause of M — of why C "indicates" the success of M — involves the fact that C carries information about, indicates, F, and the fact that F is the set of conditions on which the success of M depends. So C indicating F is involved in this *explanatory* model, but not in a causal model of how C comes to be selected. Dretske claims that C is a representation of F because this indicative relationship with F is a necessary part of a full *explanation* of why C has been recruited as a cause of M.

Unfortunately, such adversion to explanation as constituting the ontology of representation is an adversion to an observer — an explaining observer in this case — and, as such, fails again to be naturalistic. Dretske's model, then, fails to be naturalistic both because it is causally epiphenomenal in its dependence on the proper learning history, and, therefore, its inability to be defined in terms of current state, and also because it is a model of representational ascription-in-explanation, not of the metaphysics of representation per se. It is, in fact, a direct example of construing what the interactive model would take as a differentiation as being more than that, as being a representation of that which has been differentiated.

Note also that neither Dretske's nor Millikan's etiological model offer an account of the possibility of system detectable error. In both cases, not only is it not open to the system to determine its own representational content, to determine what its own representations are supposed to represent, it is also required that that content, once obtained, be compared to the current target. But, as for Fodor's model, representing the current target is the original problem of representation that was to be accounted for. In effect, this argument against system detectable error is the classical radical skeptical argument that we cannot check our representations because to do so is simply to use them again, producing a circular "check".

It should be noted that standard models of representation in Artificial Intelligence and Cognitive Science fare no better — worse, in fact. The Symbol System Hypothesis construes representation as isomorphism (Newell, 1980; Vera & Simon, 1993), but does not address any of the fatal issues that such a model encounters, such as the possibility of error (Bickhard & Terveen, 1995). Connectionism posits representations constituted as distributions of activations across a space of (output) units, but the relationship between such an activation vector and what it is supposed to represent is merely one of (at best) informational correspondence, with all of its attendant problems. In general, such vexing issues as accounting for the possibility of error, or system detectable error, are simply not addressed (Bickhard & Terveen, 1995).

I have mentioned several important problems — fatal problems, I argue — for standard models of representation: accounting for the possibility of error, accounting for the possibility of system detectable error; etiological epiphenomenalism; and dependence on unnaturalizable observers. One of the deepest errors, however, is that these models cannot account for *emergent* representation. Whether via informational correspondences or structural isomorphisms or proper functions, there is no account of emergent content for the relevant system itself — content that is in some relevant sense possessed by the organism, not by an observer of that organism; content that might permit the organism to be in error, and, perhaps, to detect that error. These models provide various ways of specifying what the content ought to be, but it is in every case a specification that is available at best to an observer analyzing the organism, not to the organism itself. There is no account of how representation could emerge — in terrestrial evolution, for example — in and for organisms themselves, as causally efficacious for those organisms, with no observers around.

The interactive model bears some interesting resemblances to — and differences from — Dretske's model that are worth commenting on. Consider the causal, not the explanatory based, version of Dretske's model. An internal state is recruited as a "cause" of a behavior because it indicates the success of that behavior. Remove the learning history considerations, and consider only the "indication of success of the behavior" relationship. That indication is internal to the system, and will involve implicitly defined properties that would support the success of that behavior. Dretske has the differentiating *state* representing such a set of properties, in virtue of its having been recruited.

The interactive model points out that it is the indication relationship itself — the internal, functional, system-accessible, indication of the potential success of M, not the external, system-inaccessible, indication of F — regardless of the history of its creation or learning, that presupposes those conditions of success. That indication presupposes the conditions of success for the behavior precisely in indicating the success of that behavior. So, it is the indication (of M), not the differentiating state upon which such an indication might be based, that represents the environmental properties, and it represents them implicitly, not explicitly as Dretske would have it. And, further, the learning history, or any other history, is irrelevant. It is not irrelevant to understanding how or why such an organization of differentiations and indications was created, but it is irrelevant to what is presupposed, and thus implicitly defined, by the predication involved in that indication — no matter what its history might be. Thus, that history is irrelevant to whether that indication constitutes a representation, and it is irrelevant to what its content might be. Still further, Dretske's model is state and action based, while the interactive model focuses on the ubiquity and greater power of interactions, both for differentiations, and as "behaviors", and Dretske is focused on a "cause" of behavior, while, as elaborated below, the interactive model involves a more abstract relationship of "indicating", rather than "causing", interactions. An *inter*action, for example, might detect an F that that organism could not detect with passive input processing; or the interaction might *create* F, something beyond Dretske's model; or the further interaction, M, that is indicated might itself, if undertaken, detect or create still further Fs. That is, it might indicate still further Ms — something completely beyond the ken of Dretske's model. The deepest contrast, however, is that Dretske's model is a spectator, backward-in-time looking model, while the interactive model construes representation in a future oriented manner, as emergent in the system's interactions, not just in its passive processing of inputs (Hookway, 1985; Rosenthal, 1983; Smith, 1987; Tiles, 1990).

More Complex Representation. A far-from-equilibrium system that is recursively self-maintenant can differentiate environmental conditions that serve to indicate the appropriateness of particular further interactions for maintaining relevant far-from-equilibrium conditions. If there are no other competing indications, that "indication" can be a kind of direct triggering of the relevant interaction — a causing of that interaction. But organisms are not always so simple nor environments so accommodating that there is only one candidate for the next interaction.

If a particular differentiation can serve to evoke indications for multiple further interactions, then some additional process must select among them, and must do so on the basis of some relevant information. A simple process for such selection would be to select among the indicated potentialities that interaction whose outcomes best satisfy, or best further, current system goals. Such an elaboration of the simple model requires an account of goals, and an account of how indicated potential interactions could be selected.

I will not elaborate the details of these extensions of the model here, but will touch upon two issues in order to indicate the nature of the corresponding extensions. First, if the function of "goal" in this model *necessitated* that goal conditions be represented, then there would be a circularity — (goal) representation being used in the model of interactive representation. Goals certainly can make use of representations, but it is not required that goal conditions be represented. In particular, goal conditions can be differentiated without being represented. What is required is that the transfers of control among the differentiation testing of goal conditions and the processes of selecting and engaging in subordinate interactions have to correct organization. That organization will be something like the TOTE structure (Miller, Galanter, Pribram, 1960) in which a Test is performed — a differentiating process — and, under some conditions, control is transferred out of the subsystem while under other conditions control will be transferred to some interactive subsystem, perhaps complex, which has some chance of furthering the goal, and which transfers control back to the Test when it is completed for another Test of the goal conditions. Thus, Test-Operate-Test-Exit, or TOTE. Obviously, this is much too simple, but my current point is simply to outline how the issue can be pursued without invoking a circularity in the model of representation.

Second, on what information is a selection of next interaction made? Again, there is more to this issue than I will address here, but one part of an answer to this question is for interactions that are indicated as potential to be associated with indications of expectable internal interaction outcomes — interaction final states — should those interactions be engaged in and be successfully executed. Those final states provide the basic information for selecting next interactions. They are the outcomes for the sake of which the selections can be made. Note further that such indicated interaction final states do not themselves require representation, thus a circularity, because they are internal, and therefore can be *functionally* indicated and tested. Such indicated outcomes also provide a direct check of the truth value of the indication: if one of the indicated outcomes is reached, then the indication was true, while if none of them are reached, then the indication was false. Such error information, in turn, can be used to guide further interaction (perhaps in a TOTE organization, or not) or to guide learning processes. Note that models that make system detectable error impossible — Fodor, Dretske, Millikan, etc. — cannot account for the very possibility of error-guided behavior or learning.

With indications of both interactions and their outcomes, not only is it possible for such indications to branch, so that multiple potentialities are indicated under a given differentiated condition, it is also possible for them to iterate. If Q is indicated as possible when A is differentiated, perhaps X and Y are indicated as potentialities should Q be in fact engaged in and its expected outcomes obtained. And perhaps X and Y each serve as the conditions for still further indications of potentiality, and so on. With branching and iterating of conditional interaction indications, they can form potentially complex webs of conditional interaction potentialities. Such webs form the basis for the further capabilities of the interactive model to account for representations such as of objects and abstractions such as numbers (Bickhard, 1980, 1993; Bickhard & Richie, 1983; Bickhard & Terveen, 1995; Campbell & Bickhard, 1986).

In this way, representation emerges as the general evolutionary solution to action selection and action evaluation, and representational capabilities become increasingly complex with more complex central nervous systems till object representation, and then abstract representation, becomes possible. There is no aporia in this model of how to get from primitive representation to complex representation or to abstract representation or to language (Bickhard, 1980, 1993, 1998b; Bickhard & Terveen, 1995).

Conclusions

The grounds of cognition are adaptive far-from-equilibrium autonomy — recursively self-maintenant autonomy — not symbol processing nor connectionist input processing. The foundations of cognition are not akin to the computer foundations of program execution, nor to passive connectionist activation vectors.

Function emerges in a particular kind of autonomous far-from-equilibrium system — a self-maintenant system. Representation emerges as a particular kind of function, indications of potential interactions, in a more adaptive kind of autonomous far-from-equilibrium system — a recursively self-maintenant system. Representational organization can become more complex via the branching and iterating, perhaps into complex webs, of conditional indications of interactive potentiality. Representation, then, emerges in the progressive evolutionary emergence of higher levels and more adaptive levels of far-from-equilibrium system autonomy.

The interactive model accounts naturally for the emergence of representation as the solution to problems of action selection and action evaluation. The interactive model is of a naturalistic emergence, a fully causally efficacious emergence — not an epiphenomenal "emergence". It is definable in terms of current state, and is not dependent on observer ascriptions or explanations or analyses. Interactive representation makes the possibility of error, and the possibility of system detectable error, easy to account for — and, therefore, makes it possible to account for error-guided behavior and error-guided learning.

Function and representation are natural manifestations of biological — or, more generally, far-from-equilibrium — autonomy. Both the understanding of natural cognitive systems and the design of artificial cognitive systems will have to accommodate these conditions for emergence.

References

Aitchison, I. J. R. (1985). Nothing's Plenty: The vacuum in modern quantum field theory. *Contemporary Physics*, 26(4), 333-391.

Aitchison, I. J. R., Hey, A. J. G. (1989). *Gauge Theories in Particle Physics.* Bristol, England: Adam Hilger.

Allen, C., Bekoff, M. (1995). Biological Function, Adaptation, and Natural

Design. *Philosophy of Science*, 62, 609-622.

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

Bickhard, M. H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, 285-333.

Bickhard, M. H. (1998). A Process Model of the Emergence of Representation. In G. L. Farre, T. Oksala (Eds.) *Emergence, Complexity, Hierarchy, Organization*, Selected and Edited Papers from the *ECHO III Conference. Acta Polytechnica Scandinavica,* Mathematics, Computing and Management in Engineering Series No. 91, Espoo, Finland, August 3 - 7, 1998, 263-270.

Bickhard, M. H. (1998b). Levels of Representationality. *Journal of Experimental and Theoretical Artificial Intelligence*, 10(2), 179-215.

Bickhard, M. H. with Campbell, Donald T. (2000). Emergence. In P. B. Andersen, N. O. Finnemann, C. Emmeche, & P. V. Christiansen (Eds.) *Emergence and Downward Causation*. (322-348). Aarhus, Denmark: U. of Aarhus Press.

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

Bickhard, M. H., Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science — Impasse and Solution*. Amsterdam: Elsevier Scientific.

Brown, H. R., & Harré, R. (1988). *Philosophical foundations of quantum field theory*. Oxford: Oxford University Press.

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). Hillsdale, NJ: Erlbaum.

Campbell, R. L., Bickhard, M. H. (1986). *Knowing Levels and Developmental Stages. Contributions to Human Development*. Basel, Switzerland: Karger.

Christensen, W. D., Bickhard, M. H. (manuscript, 2001). The Dynamic Emergence of Normative Function.

Christensen, W. D., Collier, J. D., Hooker, C. A. (manuscript, 1999). Autonomy, Adaptiveness, Anticipation: Towards autonomy-theoretic foundations for life and intelligence in complex adaptive self-organising systems.

Christensen, W. D., Hooker, C. A. (1998). From Cell to Scientist: Toward an organisational theory of life and mind. In J. Bigelow (Ed.) *Our Cultural Heritage*. (275-326). Australian Academy of Humanities, University House, Canberra, Australia.

Christensen, W. D., Hooker, C. A. (2001). Autonomy and the Emergence of Intelligence: Organised interactive construction. *Communication and Cognition,* this issue.

Christensen, W. D., Hooker, C. A. (in press). The ascent of endogenous control: Autonomy-theoretic foundations for biological organisation and evolutionary epistemology. In W. Callebaut and K. Stotz (eds.) *Bioepistemology and the challenge of development and sociality.* Cambridge, MA: MIT Press.

Clark, A. (1997). Being There. MIT/Bradford.

Collier, J. D. (1988). Supervenience and Reduction in Biological Hierarchies. In M. Matthen, B. Linsky (Eds.) *Philosophy and Biology*: Supplementary Volume 14 of the *Canadian Journal of Philosophy.* (209-234). University of Calgary Press.

Collier, J. D. (1999). Autonomy in Anticipatory Systems: Significance for Functionality, Intentionality, and Meaning. In D. M. Dubois (Ed.) *Proceedings of CASYS'98, The Second International Conference on Computing Anticipatory Systems.* New York: Springer-Verlag.

Cummins, R. (1996). *Representations, Targets, and Attitudes.* MIT.

Davies, P. C. W. (1984). Particles Do Not Exist. In S. M. Christensen (Ed.) *Quantum Theory of Gravity.* (66-77). Adam Hilger.

Fodor, J. A. (1987). *Psychosemantics.* Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A Theory of Content.* Cambridge, MA: MIT Press.

Fodor, J. A. (1991). Replies. In B. Loewer, G. Rey (Eds.) *Meaning in Mind: Fodor and his critics.* (255-319). Oxford: Blackwell.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science went wrong.* Oxford.

Godfrey-Smith, P. (1994). A Modern History Theory of Functions. Nous, 28(3), 344-362.

Hookway, C. (1985). *Peirce.* London: Routledge.

Kemp, S. (1998). Medieval Theories of Mental Representation. *History of Psychology,* 1(4), 275-288.

Kim, J. (1993). *Supervenience and Mind.* Cambridge University Press.

Kim, J. (1997). What is the Problem of Mental Causation? In Chiara, M. L. D., Doets, K., Mundici, D., van Benthem, J. (Eds.) *Structures and Norms in Science.* (319-329). Dordrecht: Kluwer Academic.

Kim, J. (1998). *Mind in a Physical World.* MIT.

Levine, A., Bickhard, M. H. (1999). Concepts: Where Fodor Went Wrong. *Philosophical Psychology,* 12(1), 5-23.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior.* New York: Holt, Reinhart, and Winston.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories.* Cambridge, MA: MIT Press.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice.* Cambridge, MA: MIT Press.

Moreno, A., Ruiz-Mirazo, K. (1999). Metabolism and the Problem of its Universalization. *BioSystems*, 49, 45-61.

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science*, 4, 135-183.

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). La Salle, IL: Monist.

Ryder, L. H. (1985). *Quantum Field Theory*. Cambridge.

Sciama, D. W. (1991). The Physical Significance of the Vacuum State of a Quantum Field. In S. Saunders, H. R. Brown (Eds.) *The Philosophy of Vacuum*. (137-158) Oxford: Clarendon.

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.

Tiles, J. E. (1990). *Dewey*. Routledge.

Vera, A. H., Simon, H. A. (1993). Situated action: A symbolic interpretation. *Cognitive Science*, 17(1), 7-48.

Weinberg, S. (1977). The Search for Unity, Notes for a History of Quantum Field Theory. *Daedalus*, 106(4), 17-35.

Weinberg, S. (1995). The Quantum Theory of Fields. Vol. 1. Foundations. Cambridge.

Notes

[1] See Christensen & Hooker (2001) for a comparison of this notion of autonomy with that of autopoiesis. Most fundamentally, autopoiesis emphasizes *independence* from the environment, while autonomy emphasizes the ability to *make use* of the environment. This makes autopoiesis a limiting framework within which to try to model higher order adaptive interrelationships between system and environment, such as intelligence, representation, and cognition.

[2] Note that the etiological approach takes *having a function* as primary, and explicates *serving a function* in terms of *having a function*, the reverse of the far-from-equilibrium approach outlined above. The issue of what type a system or subsystem belongs to, therefore, is relevant at the first step of the explication.

[3] Note that content emerges here in a form of anticipative, interactive, dynamic presupposition. Dynamic presupposition, then, can take related but nevertheless differing forms, with function and representation being two of them.

[4] The inferred properties would be more complex than that because the differentiation involved is not that precise. For example, the bacterium will continue to swim up a saccharin gradient too.