Representation, evolution, and embodiment

Michael L. Anderson
Institute for Advanced Computer Studies
University of Maryland

Abstract. As part of the ongoing attempt to fully naturalize the concept of human being—and, more specifically, to re-center it around the notion of agency—this essay discusses an approach to defining the content of representations in terms ultimately derived from their central, evolved function of providing guidance for action. This 'guidance theory' of representation is discussed in the context of, and evaluated with respect to, two other biologically inspired theories of representation: Dan Lloyd's dialectical theory of representation and Ruth Millikan's biosemantics.

Keywords. action, intentionality, evolution, representation, mental content

1. Introduction

Recent work in theoretical cognitive science can be fruitfully characterized as part of the ongoing attempt to come to grips with the very idea of *homo sapiens*—an intelligent, biological entity—and its signature contribution is the emergence of a philosophical anthropology which, *contra* Descartes and his thinking thing, instead puts doing at the center of human being.[1] This shift in our understanding of human nature is owed proximally to the work of such figures as Heidegger and Merleau-Ponty, but has clear precursors in, for instance, Hegel, and, more specifically, Marxist interpreters of Hegel such a Kojève. Naturally, Darwin must be considered as important as any philosopher, and Aristotle, too, is an important resource whose depths have been insufficiently plumbed.

What has thus been put (back) into play is a vision of human being and its place in nature that takes as its foundation not the abstract contents of Mind (belief, desire, intention), but rather action and interaction, behavior situated in a circumstance. This alternate view begins from the supposition that what is fundamental to a being is *agency*, and that as an agent it is generally integrated with, appropriate to, and selected for its environment. An agent is what it is because of its interaction with that environment, in both its individual and its evolutionary past. For those who find this latter way of thinking compelling (or who are at least interested enough to see where such a fundamental shift in thinking might lead) the broad challenge is to determine whether (and how) it is possible to build sound versions of the central concepts of

[1] I describe this shift in more detail in (Anderson, 2003a).

cognitive science—cognition, computation, perception, representation, and the like—on such a foundation as this.

I begin with such general considerations because it is important to remember that accounts of these subjects, insofar as they are driven by shared fundamental assumptions, are importantly interrelated, although they are generally treated separately. Consider, for instance, the question of the nature of perception. In approaching this question, the Cartesian first asks how it is that the features or elements of the outside world can be captured and re-presented inside the organism. Note that this simple re-formulation of the central question of perception, featuring the contrast between "inner" and "outer", and introducing the notion of representation in this context, points us in the direction of the familiar and intractable anxieties bequeathed us by Modern Philosophy: how is it possible to relate the accessible, inner given to the outer reality? Likewise, how can one adjudicate between the well-known and easily accessible self and the mysterious and potentially hostile other? Thus has contemporary philosophy been dominated by such versions of these questions as how to determine the accuracy of representations and the truth of the statements they compose, how such things as language and communication are even possible, and whether ethics (or altruism or any apparently self-sacrificing behavior) is rational.

In contrast, what has been variously called the active, enactive,[2] interactive,[3] embodied, or action-selection[4] view begins with the assumption that perception is first and foremost an organism's means for negotiating its environment. This suggests at least two things: first, that perception is a tool of exploration, and second that it is intimately bound up with, is primarily fitted to the service of, action. Perception is not the passive reception of abstract qualities from the environment, but is itself active, often highly selective and goal-directed, designed to provide guidance for the current (or ongoing background) purposes of the agent. The primary task of perception, then, is not the construction of objective models (which is not to say that organisms never build models), but the detection of opportunities for action, a notion that recalls the familiar phenomenological claim that the perceptual field is always an action-field, that the perceived world

[2] (Varela, Thomson and Rosch, 1990).
[3] (Bickhard, 1993; 1999)
[4] (Franklin, 1995).

is always known in terms directly related to an agent's current behavioral options. To put it in terms of affordances (Gibson 1966; 1977), the perceived availability of things to certain interventions: the world is *seen as* a continuous series of invitations to action.

As with perception, so with representation. In what follows I will describe an approach to defining representational content in terms ultimately derived from the central, evolved function of representations to provide guidance for action. I will then contrast this 'guidance theory' of representation with two other biologically inspired theories of representation: Dan Lloyd's dialectical theory of representation and Ruth Millikan's biosemantics. The essay functions as a sequel and supplement to (Rosenberg and Anderson 2004; forthcoming), but is in no way a replacement for the detailed, formal account of the guidance theory offered there.

2. Representation and evolution

What is representation? More to the point, how should we approach the question of understanding representation in animals (including humans)? Broadly speaking, a representation is any entity (object, property, state, event) that designates, denotes, or stands in for another. Although there are likely to be many conditions that a given entity must (or could) meet to fit this broad definition, in the case of natural, biological representation, the ultimate question any theory of representation must answer is: what is the nature of the relation between a specific event or state **R** in a living system, to some other specific event, state, object or property **E** that permits one to say that **R** represents **E**? A typical account centers on the notion of an internal state reliably triggered by (or in some other, specifiable, way *dependent on*) a certain class of stimuli; let's call accounts of this sort *detection* theories of representation. To see how detection theories work, in the context of a more concrete example, let's consider a hypothetical animal that bears a resemblance to the common frog *R. pipiens*, in an environment that bears a resemblance to Earth.[5] Let us suppose further that this animal reliably generates two specific, neural signatures—two specific patterns of excitation in a certain region of its brain—one of which is dependent on the presence of moths in its visual

[5] For some of the fascinating details of the representational mechanisms in the real *R. pipiens* see, e.g. (Lettvin *et al*., 1959; Ingle, 1973; 1980; 1982).

field, and the other on the presence of gnats. It is of course the case that the retinal stimulations will differ here, but the supposition is that, in addition, there are differentiable neural signatures somewhere further down the processing stream. Detection theories of representation are those ultimately grounded in the following sort of story: the occurrence of a given event of the animal (a particular neural signature) is a representation of gnats just in case it depends upon the presence and perception (the *detection*) of gnats in the environment. The gnat-representation is the inner state reliably correlated with sensing gnats (in general, and in any particular case), an inner state that perhaps also co-varies with, is isomorphic to, or carries information about the gnat (and/or about gnats more generally). There are many different detection theories of representation, for instance, *causal theories* holding that a given representation **R** is about **E** just in case it has a certain specified set of causal relations to **E**, for instance, that perceiving an instance of **E** will cause one to represent with **R** (Fodor, 1981; 1987; Lloyd, 1990; Millikan, 1984), and *information content theories* holding that a given representation is about that object from which the information it contains in fact derived (Dretske, 1981; 1986; 1988). At best, however, detection theories tell only half the story. For to understand representation, it is crucial to understand not just what the environment does to the animal, but what the animal does with the representation.[6] This aspect of representing is left out of detection theories, but it is crucial to a full understanding of what representation in animals actually is.

Note that in the most general sense what has been left out of (or pushed to the background in) detection theories of representation is the *agency* of the animal. This puts these theories rather directly at odds with the general approach to mind that naturally emerges from a more evolutionarily grounded perspective. From the standpoint of evolutionary theory, the various heritable structures of the organism should be understood in terms of their functional contributions to genetic survival, the two main aspects of which are general contributions to the physical viability and longevity of the organism, and more specific contributions to reproductive success itself. The central nervous system (CNS), of course, contributes in both ways, but primarily in virtue of its central, evolved function of choosing, guiding, and monitoring the

[6] Arguments for this thesis can also be found in (Lloyd, 1990) and (Millikan, 1993); their theories of representation will be briefly discussed, below.

organism's *behavior.*[7] Thus, from the evolutionary perspective, the CNS should be understood primarily as the behavioral control system of an environmentally integrated agent, and representation as one (extremely important) element in an overall solution to successful environmental coping. When agency is brought back to the fore in this way, certain obvious facts about representation become more significant, and lead to a substantially revised picture of its nature and basis. First of all, consider that for an agent to *use* a representation requires both that the representation has features that are amenable to use, and that the agent has mechanisms to exploit those features. Indeed, we might go further and say that, for a given inner state to have a meaning for the agent—for it to *represent* something for that agent—the agent must have mechanisms that actually *do* exploit those features. For consider our hypothetical frog. It is certainly true that the frog's brain, as described, is a reliable detector of (and differentiates between) gnats and moths. The brain *could be used* to differentially indicate, and therefore in the broadest possible sense to represent, gnats and moths. Should we wish to use the brain in this way, the two neural signatures could represent gnats and moths *for us.*[8] But what warrants the claim that the agent whose brain this is differentially represents gnats and moths? I should like to say that the claim is warranted if and only if each inner state (neural signature) *does something different for* the agent, if the difference that exists in the neural signatures is in some way exploited by the agent.

Thus, for instance, if the agent exploits the difference in neural signatures to trigger different feeding strategies,[9] or if each inner state allows different inferential (or proto-inferential) moves—that is, it triggers different *further* representations—then it seems warranted to say that the differences are indeed significant for the agent. This fact—and not the mere existence of differentiable inner states—is the necessary grounding of the claim that these inner states represent. For consider the case where the agent does *not* in any way exploit the different features of the inner states; its behavior, and the effect on its mental state, is in

[7] Smarter, more flexible creatures may, of course, enhance their survival rate (and therefore their overall chance of reproducing) in virtue of enhanced environmental coping skills, but intelligence can also play a more direct role in mate attraction (wit as aphrodisiac). Also fundamental is the role of the CNS in general, and the brain stem in particular, in homeostasis, but this is not relevant to the issue of the nature of intelligence/cognition.

[8] We have our own methods of gnat and fly detection, and so have no use for a frog's brain in this context. However, the use of biological detectors has a long history: trail-sniffing dogs, gas-detecting canaries, etc.

[9] *R. pipiens* captures small dark things with its tongue, but displays different behaviors in the presence of different prey (Deban *et al.* 2001).

each case identical. This might be the case, for instance, if the frog used *only* those aspects of the neural signatures that specified location, to aim its body and tongue. In such a case, it would seem untenable to claim that the frog differentially represented gnats and moths; at best it would represent the disjunction <gnat or moth>, with *both* neural signatures having this same meaning.[10] And what if the frog in no way exploits (is perhaps incapable of exploiting) the neural signatures *at all*? If the inner states which are in point of fact reliably correlated with (caused by, etc.) gnats and moths make absolutely no behavioral or cognitive difference to the animal, it not only seems unwarranted to claim that it *represents* gnats and moths, it sounds strange even to say that it *detects* them.[11]

The general point is this: to specify the meaning to an agent of a given inner state—indeed, to know if it is a representation for that agent at all—it is necessary to determine its actual value to, and use by, the agent doing the representing. In fact, as was demonstrated by Rosenberg and Anderson (2004; forthcoming), it is not only *necessary* to do this, it can be *sufficient*. That is, an approach to understanding representational content that focuses on how a given inner state is used by an agent to guide its behavior is sufficient not only to show what the inner state means for the agent, but can also do the work of specifying what in the environment it stands in for. Adopting the general definition of representation given by the guidance theory to the specifically biological case: a state **R** in agent **A** represents entity **E** for **A** in circumstances **C** just in case **A** has an enduring conscious preference or conditioned reflex to use **R** to guide its behavior with respect to **E** in **C**. Obviously a great deal rests on that last clause, *with respect to **E***, and while it is not possible to go into detail here the general idea is relatively straightforward and intuitive. For instance, a motor action **M** is taken *with respect to* an entity **E** when **E** is the focus of agent **A**'s effort to change the world by **M**. In the case where the action **M** involves setting up feedback loops with the environment to monitor the action's outcome, the focus of an action can be further identified in terms of the ultimate entity

[10] Note that none of this implies that the frog represents moths *as* moths and gnats *as* gnats, as this may require conceptual or other abilities not here supposed. Indeed, part of the point is to emphasize that the actual content of a representation must be understood in terms of what the agent does with that representation, which would include thinking or reasoning with it. The frog as described may represent gnats non-conceptually (Chrisley, 1995) as tongue-prehendible-prey and moths as jaw-prehendible-prey, but only in the case where some such distinction is indeed made can one say that it differentially represents both.

[11] One would be lead to say such things as that the perceptual system detects them, but the animal does not represent or notice them, much as one might say that my eye detects photons, but I do not.

being monitored by **A** as part of efforts to track the success and effects of **M**.[12] In the case where the action does not involve feedback and monitoring, other expedients must be employed to determine the action's focus, but in each case the idea is ultimately to identify **E** in terms of non-representational facts about how the agent *works* and what it *does* when using **R** in **C**.[13] The interested reader is encouraged to consult (Rosenberg and Anderson 2004; forthcoming) for a fuller treatment.

3. Representation and biology

To better see the significance of this move from *detection* to *guidance* as providing the fundamental grounding for a theory of intentional, representational content, it is perhaps useful to contrast the guidance theory with two other biologically-inspired theories of representation, both of which accept, with different results, the central point that any viable theory of representation must address not just representation *production*, but also representation *consumption*: Dan Lloyd's dialectical theory of representation and Ruth Millikan's evolutionary biosemantics.

3.1 The dialectical theory of representation

Lloyd approaches the problem of representation primarily by asking what it would take to build—or evolve—a simple representing system in a given environment. This puts him fairly squarely in the same bottom-up, biologically oriented camp that generated the guidance theory, and the perspective does indeed seem to exert an influence on his intuitions. Thus, for instance, he argues, as we did above, that whatever else representations are, they must play some sort of (cognitive) role in explanations of behavior if they are to *be* representations. The insight, in turn, naturally leads him to consider the viability of various internalist

[12] Such monitoring can be direct or indirect. One might monitor the focus directly, as when our imagined frog monitors a gnat to determine when its feeding action, guided by **R**, has succeeded; or indirectly, as when an engineer moves a knob to increase the pressure in a boiler, and monitors a gauge because he takes it to provide feedback about the status of the pressure in the boiler.

[13] One immediate benefit of the guidance theory, which distinguishes it from detection theories of representation, is that for a token to be used to guide action does not require the current *presence* of the represented entity. That representations can be legitimately used, and legitimately triggered, in situations where the represented entity is not present (e.g. in planning for future action) is one of the great benefits of using representations in intelligent systems. Any theory of representation, then, ought to easily account for this feature, which might be called the *potential decoupling* of the representation from its object. See (Rosenberg and Anderson, forthcoming) for further discussion of this point.

theories of representation, which argue that representational content can be fixed in terms of "relations among events internal to the organism or representing system." (Lloyd, 1989:23) Conceptual role theories, for instance, try to analyze meaning in terms of the role played by the concept in inferential and other conceptual/cognitive processes: roughly speaking, the representation **R** is about **E** just in case it is used to make warranted inferences about **E** (Harman, 1982; 1987). Likewise, according to functional role semantics (Block, 1986), the meaning of a particular representation in a system is given by the functional role it plays in that system. However, although Lloyd seems attracted to the notion that the functional (or cognitive) role of a representation is surely relevant to its content, and is apparently willing to entertain the idea that, given a representational system, the content of a given item in that system can be given by its role in that system, in the end Lloyd rejects internalist theories because they cannot answer the fundamental question: what makes such a system representational in the first place.

To motivate this conclusion, Lloyd argues, first, that defenders of internalist theories must ultimately fall back on the claim that the individual tokens represent because of their place and role in a representing system, but that *entire systems* represent because they exhibit some kind of isomorphism to the structure or system being represented (*see also* Fodor, 1987). However, citing the familiar Twin Earth-style examples, Lloyd argues that isomorphism cannot do the work required of it here. Thus, for instance, suppose that I am imagining my backyard, with its particular arrangement of deck furniture and plants. What makes this image an image of my backyard? Neither resemblance nor some more general isomorphism will do, here, for suppose an exact duplicate of my backyard exists somewhere. By hypothesis, all the isomorphisms that exist between the image and my backyard will hold between the image and this *other* backyard. Yet, surely (the argument goes) the backyard I am imagining is *mine*. Likewise, consider two identical images, one a photograph depicting my mother, and the other the chance result of a boiler explosion. Surely only the photograph is a picture *of* my mother, even if both resemble her.[14] Lloyd suggests, following the standard line, that to preserve our intuitions here we must turn to externalist theories, which define representation "as a relation of probabilistic dependency between a representing event and the event it represents." (Lloyd,

[14] If you don't like these examples, consider a similar case of two *really identical* twins; a picture which resembles both is only a picture of one of them.

1989:23) I imagine *my* backyard, because that is the one on which my image in fact depends; likewise, only one picture is *of* my mother, because only one bears the right relations of dependence to her.

Now, it is far from clear to me that these sorts of arguments should carry much weight in discussions of the nature and origins of biological representation. Biological systems, after all, did not develop so as to able to differentiate between one backyard, and an identical one on Twin Earth, and so perhaps from the standpoint of the actual content of biological representations, there *is* no way to differentiate these. That is, even if a theory of representation rooted in isomorphism *does* lead to the conclusion that content is ambiguous or underdetermined, this ambiguity (and therefore the theory) cannot necessarily be ruled out on *biological* grounds. Our knowledge that, in point of fact, a representation was formed in one place rather than another (and so in fact has a certain set of dependency relations to that place) may give *us* grounds for the claim that it represents *that particular* place, but that does not necessarily license one to import the considerations *we* would use to make this determination into the content of the representation as it exists in the agent under consideration.[15] Especially if we have accepted the general notion (advanced above) that understanding the content of a representation is fundamentally a matter of understanding its meaning to (that is, its role in and use for) the agent doing the representing, such arguments should at the very least raise eyebrows, if not alarm bells. Furthermore (and more importantly), these arguments tend to rely on examples like photographs, portraits, and memories, which appear to have features that do not carry over to the general case of representation. Although it is true that part of coming to know the representational content of a photograph or portrait *is* coming to know its causal history, this is because part of what it is to *be* a photograph or a portrait is to have a specific causal history that determines its representational content. It seems perfectly reasonable in this case to imagine having to say things like: we *thought* this was a picture of Pharaoh Ramses, but it turns out it is a picture of Amenhotep.[16] Likewise, the contents of episodic memories are determined at least in part by particular facts about causal history. Here again, this is because, insofar as an episodic memory is a record of a past event, *which* past event was in fact recorded is naturally quite important to its content. We might call causal histories with this sort of relation to content *continuing*

[15] Nevertheless, this line of thinking leads directly to the so-called "long-armed" single-factor versions of FRS (Harman, 1987), and to two-factor theories (White, 1982) .

[16] This can be the case whether or not Pharaoh Ramses resembled Amenhotep (so far as we know, he didn't).

determinants. However, consider in contrast such things as semantic memory, procedural memory, and "muscle memory". Here it seems that what is important is what might be called present content (the fact that was learned) or present function (the skill that was gained), and not the details of how that content or function was in fact acquired. It is of course always true (indeed, almost tautological) that the causal history of a state determines (in the sense of *causes*) its content, but here it is the *effect* of that history which matters to (perhaps *is*) the content, and not the history itself. We might call such a history a *transient determinant* of content. Unlike in the case of photographs, here it makes little sense to say: I *thought* I knew how to play piano (for I evidently have the skills of piano-playing), but, having discovered more about the causal history of my skill acquisition it turns out that I actually know how to play the tuba.[17] Thus, although representations, concepts, and the like will all have been formed by some particular causal process with a specific history, that history may well be a transient determinant, and thus need not be considered a *part* or *continuing determinant* of present content, as it apparently is for a photograph.

For consider the case where a series of random events leaves in its wake a cognitive structure (let's call it a "concept") that allows an agent (and is in fact used by that agent) to choose appropriate actions in its dealings with water bugs. This structure has just the right connections with the agent's perceptual, inferential, and other action-producing components so that the role it plays for the agent is to guide its actions with respect to water bugs.[18] What, exactly, prevents the conclusion that this "concept" represents water bugs? For the guidance theory, nothing does; the guidance theory treats causal history in the general case as a transient determinant of content. And surely we don't want to argue along the following lines: while it seems that this structure is *used by the agent* to represent water bugs, it doesn't *really* represent water bugs, owing to its aberrant causal history. Must we accept the conclusion that, if a mental event or state represents anything, it can *only* represent its causes? And in this case, since the state does *nothing* for

[17] Being clever folks, we can surely construct the appropriate thought experiment, wherein as a result of some strange and highly non-standard circumstances, a great deal of practicing on the tuba resulted in the skills for playing the piano (one imagines the poor student in this scenario getting *worse and worse* at tuba-playing). For historical accuracy, the student *might* say something like: I learned to play the piano *by* playing the tuba. However, one does not, in virtue of altering the causal history of a skill, thereby change its content. Whatever its causal history, it remains the ability to play the piano.

[18] Indeed, if we want to push the example, we can stipulate that the agent *had* just such a structure, formed by the normal methods for generating such a structure (which presumably involve repeated encounters with waterbugs), but this original structure was wiped out by a lightning strike. A half second later, however, a new, identical structure was formed as the result of a freak burst of gamma radiation.

the agent having anything to do with those causes (and we are still entertaining the notion that a representation must have *some* sort of function), must we therefore conclude it represents *nothing*? The conclusion does not seem attractive.[19]

The case appears likewise for an agent whose cognitive mechanisms were formed in the normal way in one environment, but who is suddenly transported to a Twin world that is, by all methods of determination available to the agent, exactly similar to (but is, if you like, in some important but hidden way different from) its original environment. If the agent uses the cognitive structures formed on Earth to guide its actions with respect to the Twin water bugs, here in its new Twin environment, it seems pretty reasonable to conclude that the structures represent Twin water bugs. The Twin environment is, after all, its *actual* environment, and the Twin water bugs are the actual bugs it is dealing with. Here again, it seems perverse to insist that the agent is *in fact* representing Earth water bugs (because the causal history of the structure connects it to those), or perhaps nothing at all (since there is nothing in its current environment with which it has the right history of causal interaction[20]). In trying to understand biological, mental representations, a certain saying about ducks comes to mind. For the guidance theory, at least, what matters in the general case of content is not the details of how a given representational inner state was in fact acquired, but its present function for the agent in guiding its actions.[21]

Still, although I think Lloyd missteps in failing to distinguish between transient and continuing determinants of content, he is hardly the only theorist to me misled in this way, and there are two more important issues to be considered. First, even granting the inability of purely internalist representational systems to anchor themselves to some particular part of some particular (non-Twin) external world,

[19] Among other problems, the position would appear to rule out the possibility of representing anything unless that thing exerted some sort of causal influence on the representation. But abstract entities, fictions, and future events do not appear to exe rt causal influence. Does this mean these entities cannot be represented?

[20] Here we might elaborate the back-story, and claim that everything was moved to Twin Earth because the actual Earth was in the way of an intergalactic highway project, and has since been destroyed.

[21] Note further that on the guidance theory representational connection does *not* depend on isomorphism—although isomorphism (along with its related properties like resemblance, co-variance and information-bearing) can play a role in explaining why a given state is *effective* in guiding action—and therefore the guidance theory is not obviously susceptible to Twin Earth style arguments in general, whatever their ultimate utility in guiding our intuitions about biological representation.

probabilistic dependency relations are not the only alternative capable of providing the required grounding. Second, if Lloyd's theory is anything to go by, adopting such dependency relations in fact makes it very hard to preserve the internalist insight that cognitive role matters to content. For in light of the considerations outlined above (among others), Lloyd adopts a three-factor theory as follows:

> A natural event r is a representation if it meets these conditions:
>
> 1. *The multiple channel condition* There is a set of at least two events, $\{v_1, v_2, \ldots, v_n\}$, such that r is dependent on the concurrent conjunction of at least two events in the set.
> 2. *The convergence condition* Events v_1 through v_n are further subject to the constraint that there is a set of single events, $\{u_1, \ldots, u_n\}$ (the single mutually effective stimuli), such that all of v_1 through v_n depend on each element of $\{u_1, \ldots, u_n\}$. The object of a representation is the element of $\{u_1, \ldots, u_n\}$ with the highest conditional probability, given r.
> 3. *The uptake condition* Event r has the capacity to cause either another representation or a salient behavioral event. (Lloyd 1989:64)

The first clause in Lloyd's theory emerges from considering the tradeoffs of reliability and error in a natural system detecting events in its environment. To neither over- nor under-detect events in the world, Lloyd argues, it is best to require at least two detectors to both indicate the same event. The events $\{v_1, v_2, \ldots, v_n\}$ are canonically events in the system (potential detections) caused by environmental stimuli. Each such event *could* be caused by one or more different stimuli; these are the mutually effective stimuli for that event. Thus, the second clause states that the actual object of the representation is the highest-probability member of the set of mutually effective stimuli. Finally, the third clause fulfils the basic requirement that for an event to be a representation, it must play some role for the system.

Because Lloyd looks at representation from the standpoint of an agent in an environment, and accepts the very basic point that a representation must be something which matters to, or has a function in, an environmentally situated agent, and yet comes to such different conclusions, his theory makes for a useful contrast with the guidance theory. On the guidance theory, action-guidance plays the role of a representation's cognitive function, allowing us to preserve the internalist insight that the functional role of a representation matters a great deal to its content, and *actions themselves* provide the necessary link to the world. According to the guidance theory action is fundamentally intentional: it is first and last a directed engagement with the world. Representations come into existence and derive their content from their role supporting the basic intentionality of action. In contrast, on Lloyd's theory, probabilistic dependency

relations (clause 2) do all the work of specifying representational content, and cognitive role is preserved only as the near-empty requirement that the representation do *something* (anything!) for the system. Thus, Lloyd has unnecessarily given up the core internalist insight, so as to be able to fix external representational content. The guidance theory does not force this choice.

The crucial moment in Lloyd's argument comes when he reaches the conclusion that functional role theories, and purely internalist theories more generally, while potentially able to specify internal content in functional terms, could not explain or ground the representational connection *per se*. The conclusion is likely sound, but I think Lloyd makes an error in framing his response. Rather than immediately turning outward to see what resources were at hand to help explain and ground this connection, he might instead have turned further inward, and asked a rather different question: granted that the representational content of an item in a representing system was a matter of its function in the system as a whole, might it not be the case that the representationality of the system itself is a matter of *its* functional role in the agent as a whole? Here again, I would like to suggest, evolutionary considerations point us in roughly the right direction, for the biological function of a representing system is, in the end, to govern behavior. And behavior—action—is the key to understanding representational content, or so I have been arguing.

3.2 Biosemantics

Returning to the notion that representation should be understood in light of biological or evolutionary function brings us to a different theory of representation, which will likewise serve as a useful contrast to the guidance theory. Millikan's approach to the mind is deeply biological and thus functional, ecological, and inevitably evolutionary. For this reason, her biosemantics is, if anything, even closer in spirit to the guidance theory than is Lloyd's dialectical theory of representation. For Millikan, psychology is "a deeply ecological science dealing with how the organism interacts with its wider environment" (Millikan, 1993:12), and "an appeal . . .to function is what is needed to fly a naturalist theory of content." (Millikan, 1993:85) Thus she writes, "cognitive systems are designed by evolution to make abstract pictures of the organism's environment and to be guided by those pictures in the production of appropriate actions," a

statement that is also a nearly perfect fit with the guidance theory.[22] (Millikan, 1993:11) Creating, and being guided by, abstract pictures is the *function* of cognitive systems, a function they have in virtue of their evolutionary history.

But while this is the context within which, and the materials out of which, Millikan fashions her theory of representation, the driving force behind the theory is the problem of error. For it is one of the key features of representations that they can be both faithful and unfaithful to the entity they purport to represent, and in her view no other theory of representation has adequately accounted for this. Millikan divides the options into three: picture theories, causal/informational theories, and "PMese theories" (so named after Sellars's term for symbolic logic). As we saw above, picture theories, which ground the representational relationship in "likeness", appear to have trouble with determinately fixing the object of representation. This can be easily put into the context of error as follows: insofar as resemblance is what *makes* a representation represent a particular entity, then any failure of resemblance, any unfaithfulness on the part of the representation, rather than simply reducing the quality of the representation (making it a poor likeness of its object), instead actually undermines its very connection to that object. It may turn out to represent something else that it resembles more closely, or nothing at all, but insofar as it does *not* resemble a given object, it therefore cannot represent it. According to the picture theory, it does not appear that representational error is possible, for a representation cannot be simultaneously unfaithful to, and yet still represent, a given entity. Thus, if we wish to acknowledge that likeness plays at least some role in representational content (and this is surely the case), it appears that the story must be supplemented somehow.

Perhaps the most common way to supplement the story (again, as we saw above in a slightly different context) is with causal dependence relations. Any given representation always has a causal history, and this history generally implicates the object of the representation. Roughly speaking, on a causal or informational theory of reference, a representation represents what causes it, or what, through some more

[22] I say nearly because, while the general sentiment is clearly compatible, the guidance theory would eschew the word "picture", and state the case somewhat more directly: "cognitive systems are designed by evolution to create inner states and structures for guiding the organism's actions in, and its interactions with, its environment."

complex relation of causal dependence, it reliably indicates. It is easy to see the attraction to this account; after all, the notion that a picture of a skunk is a picture of a skunk just in case it was *caused* by (or depended on the presence of) a skunk, has a certain intuitive appeal. But it is also easy to see that, in this simple form, causal theories fall into the same kind of problem that picture theories do. For effects always have a cause—*some* cause—and so when the effect is a representation, and cause dictates its object, then it seems that to fail to represent would be to fail to have been caused, which results in a neat little contradictory circle. Alternately, if there are uncaused effects, then representations too must sometimes be uncaused, in which case they represent nothing. Perhaps we are tempted to say: they represent their cause when they have one, and nothing when they do not, but this, of course, is just to say that representations always correctly represent, which they manifestly do not. And finally, if we grant that representations can have multiple causes (including, perhaps, no cause), then we face the question of which of these causes they in fact represent. To paraphrase Millikan's example, if we find that our representation of skunks can sometimes be triggered by the presence of cats, does that mean it is really a representation of <cats or skunks>, since this is the (strange, disjuncted) entity it reliably indicates? (Millikan, 1993:7)

What is pretty clearly called for in the case of causal theories is some account of its *proper*, *standard*, or *normal* cause. That seems right, but Millikan is quick to point out that the most obvious candidate for such an account, statistical reliability—which would make the most *common* or *probable* cause the object of the representation—is pretty unlikely to be of use here. More often than not, the beaver slaps his tail on the water, and the white-tailed deer puts hers in the air, in the *absence* of danger. Does this mean that these signals in fact indicate nothing? Beavers and deer certainly don't *act* as if they do. (Millikan, 1993:90-1) While one can certainly argue with the particular examples (perhaps these signals *don't* indicate danger, but rather: run!), the general point seems sound, and well worth consideration in this context:

> Many biological systems perform their proper functions not on the average, but just often enough. The protective coloring of the juveniles of many animal species, for example, is an adaptation passed on because *occasionally* it prevents a juvenile from being eaten, though most of the juveniles of these species get eaten anyway. Similarly, it is conceivable that the devices that fix human beliefs fix true ones not on the average, but just often enough. If the true beliefs are functional and the false beliefs are, for the most part, no worse than having an empty mind, then even very fallible belief-fixing devices might be better than no belief-fixing devices at all. . . . Perhaps, given the difficulty of designing highly accurate belief-fixing mechanisms, it is actually advantageous to fix too many beliefs, letting some of these be false, rather than fix too few beliefs. (Millikan, 1993:91)

However this particular point is received, specifying the conditions for the normal or proper cause of a representation remains very much an open project.

This brings us to the third and final of Millikan's three options for a theory of reference, PMese theories. PMese theories are, more or less, what Lloyd called internalist theories: they define the representational content of a given item in terms of the functional role that item plays in an inferential system. The meaning of a given item, therefore, is determined or defined by the inferential transforms in which it in fact participates. As Millikan points out, however, if this is the arrangement, it is unclear how to differentiate between valid and invalid inferences. Typically, one can make this distinction because the meanings of terms and rules are fixed *independently* from any attempted actual inferences. Those inferences are valid which conform to these pre-established rules and meanings. But if the meanings are defined by the actual inferences performed, then presumably the meanings will be whatever is necessary to make (will be defined as whatever, in fact, makes) these inferences valid. Considering the issue of perceptual representation in this light, if my perceptual/inferential system in fact responds to cats with "skunk", then "skunk" will *mean* cat (or perhaps, as above, <skunk or cat>) for this is the meaning required to make the transform valid. Once again, we are faced with the necessity of supplementing the story with some account of *normal* or *proper* inferential functioning.[23] Only if we have such an account in place will we be in a position to explain why, despite the fact that one sometimes thinks "skunk" when seeing a cat, "skunk" nevertheless means skunk, and "cat" means cat, and representations are sometimes mistaken. (Millikan, 1993:9)

As already noted above, Millikan is unimpressed with efforts to cash out normal or proper in terms of statistical reliability. Instead, she says, we need to return to the core idea of *biological* norms and functions. Pushing an analogy with photographs, she asks: is it not the case that cameras reliably produce likenesses because of how they are designed, and is it not the case that it is this design that dictates the conditions under which it will function properly, and thus produce true likenesses? Likewise, if we take

[23] One different way to go, roughly following Lloyd, would be to solve the problem by grounding the inferential system in the world, thus fixing the meanings of inferential items in some form of external relations.

our brains to have been designed to produce representations (Millikan, 1993:6), then this *design* is what will determine the meaning of proper function, the conditions under which it will produce true representations. Accordingly, Millikan writes:

> It is not the facts about how the system *does* operate that make it a representing system and determine what it represents. Rather, it is the facts about what it would be doing if it were operating according to biological norms. When functioning properly, a mental representation co-occurs with its represented, pictures what it represents, and (if it is of the right, rather sophisticated sort) participates in appropriate inferences. (1993:10-11)

The details of Millikan's theory are neither easy to understand, nor to recount. The basic idea, however, is clear enough.

> First, consider beavers, who splash the water smartly with their tails to signal danger. This instinctive behavior has the function of causing other beavers to take cover. The splash means danger, because only when it corresponds to danger does the instinctive response to the splash on the part of the interpreter beavers, the consumers, serve a purpose. If there is no danger present, the beavers interrupt their activity uselessly. (Millikan, 1993:90)

In other words, however often the tail splash *correctly* signals danger, it nevertheless *always* signals danger because that is the proper function of the tail splash. The reason the instinct both to produce and to react to tail splashes was selected for, and continues to be preserved, is that the tail splash occurs often enough in the presence of danger to be functional. Or, to put it slightly differently, the interpretation that gives the signal its biological function is the one that fixes its meaning.

Millikan tells a similar story about a certain species of aquatic, magnetotropic bacteria. In the Northern Hemisphere, the magnetosomes in these bacteria are oriented such that they pull the bacteria toward magnetic north, thus down, and thus toward deeper, oxygen-poor waters. Since oxygen is toxic to this species, the pull toward deeper water has a clear biological function, which leads Millikan to say that the magnetosomic pull *indicates* the location of oxygen-free water. (Millikan, 1993:92-3) This is the interpretation of the pull that accords with its biological function (and thus it means "oxygen-free water" even when, for instance, an aberrant magnetic field causes the bacteria to travel up into toxic, oxygen-rich waters).

More precisely, Millikan's biosemantics fixes content according to something like the following principle: **R** represents **E** just in case the presence of **E** is what makes sense of **R**, or is required for **R** to be functional; alternately, **R** represents **E** just in case **E** corresponds by some rule to **R**, and the absence of **E** would disrupt the function of **R**, with the further condition in each case that the function in question is a *proper function*. She writes:

> Note that the proposal is not that the content of the representation rests on the function of the representation or of the consumer, on what these do. The idea is not that there is such a thing as behaving like a representation if *X* or as being treated like a representation of *X*. The content hangs only on there being a certain condition that would be *normal* for performance of the consumer's functions, namely, that a certain correspondence relation hold between signs and the world, whatever those functions may happen to be. (Millikan, 1993: 89)

Content, then, is to be fixed not directly in terms of function, but by discovering what correspondent (and what relation of correspondence) is required for an item or mechanism to *perform* its function. In addition, the function is not just whatever the item or mechanism happens to do, but is its *proper* function.

> To put things very roughly, for an item *A* to have a function *F* as a "proper function", it is necessary (and close to sufficient) that one of these two conditions should hold. (1) *A* originated as a "reproduction" (to give one example, as a copy, or a copy of a copy) of some prior item or items that, *due* in part to a possession of the properties reproduced, have actually performed *F* in the past, and *A* exists because (causally historically because) of this or these performances. (2) *A* originated as the product of some prior device that, given its circumstances, has performance of *F* as a proper function and that, under those circumstances, normally causes *F* to be performed by *means* of producing an item like *A*. (Millikan, 1993: 13-14)

Although the general idea is, of course, to provide an analysis of proper function in terms of the history of a given item, the canonical example of such a history (and the one of greatest relevance to us here) is, of course, evolutionary history; thus, item *A* has proper function *F* if *F* is the result of, or is being currently preserved by, natural selection. In the case of the magnetotropic bateria mentioned above, the story goes like this: the magentosomes caused the bacteria to have certain behavioral tendencies—swimming down—that improved their survival because deeper water contained less oxygen (or, if you like, in the past the bacteria with magnetosomes causing this behavioral result were the bacteria that survived longer and produced more offspring, explaining the presence of the trait in current bacteria). Causing the bacteria to swim to deeper waters, then, is the proper function of the magnetosomes. Further, the environmental correspondence with the pull of the magnetosome that is required for it to *have* this function is the presence

of oxygen-free water. Thus, on the biosemantic account, the pull of the magnetosome indicates (or represents) the presence of oxygen-free water. Likewise for the tail-slapping of the beaver; it indicates danger just in case the behavior qualifies as a proper function, and the presence of danger is the environmental correspondence required for the behavior to *have* this function.

3.3 The guidance theory

We are now in a position to see some of the differences between biosemantics and the guidance theory. For while the guidance theory *also* gives the result that the tail-slap indicates the presence of danger, the basis for this claim is a bit different: the tail-slap indicates danger just in case the signal is standardly used[24] by the beavers to guide their actions with respect to danger. More specifically, in the case under consideration the tail-slap indicates danger just in case aspects of the signal are used to guide motor actions that (a) have danger as their focus[25], and/or (b) result in the setting of feedback loops with the environment in which danger is the ultimate entity being monitored, and/or (c) involve an assumption of information about danger as a motivating reason for the actions.[26] Thus, the tail-slap indicates danger on the assumption, for instance, that beavers move away from the direction of the splash, or move to the nearest shelter taking into account the location of the splash.[27] An even stronger case for the representational function and content of the signal could be made if, for instance, beavers monitor the location of the splash for *further* signs of danger, choosing or modifying their actions according to the results of that monitoring. In contrast with biosemantics, then, the guidance theory gives more importance to the current or actual role of a representation in guiding behavior, and does not explicitly consider the origin of that role. More generally, although evolutionary history can of course help determine what the function of a given structure is, and environmental correspondences of the sort Millikan leverages often play a role in the explanation of why one structure rather than another was preserved, according to the guidance theory these facts nevertheless

[24] An item is "standardly used" for *X* if the agent "has an enduring conscious preference or conditioned reflex to use" the item for *X*. (Rosenberg and Anderson 2004; forthcoming)
[25] Actions taken, for instance, to minimize, avoid, or confront danger.
[26] Where "assumption of information" and "motivating reason" are ultimately cashed out in terms of non-representational facts about what an agent does in its circumstances. See (Rosenberg and Anderson 2004; forthcoming).
[27] Such modulation of behavior in light of different aspects of the signal (e.g. its location) is something like a minimal condition for using the signal as a representation; the behavior-determining elements of the agent must exploit given features of the signal.

do not determine representational content. One might say that, whereas biosemantics treats these historical facts as *continuing determinants* of content, on the guidance theory they are *transient determinants*. What matters is what the representation does for the agent, and what the agent does with it.

The case is likewise for the magnetotropic bacteria. Although for the guidance theory the bacteria are probably too simple to represent anything[28], if one *were* to make the case that the pull of the magnetosome indicates the presence of oxygen-free water for the bacteria, the *basis* of the claim would rest neither directly on evolutionary history and the correspondences it exploited, nor on the immediate guidance provided by the pull itself. For with respect to what, it must be asked, does that pull guide the bacteria's behavior? The answer is that the bacteria's behavior seems to be immediately guided only with respect to the magnetic field itself. Unlike the beaver's actions in the case of a danger signal, the bacteria do not monitor, modulate their behavior with respect to, or otherwise mark the presence of oxygen-free water. The magnetosome simply pulls the bacteria deeper, by simple and direct causal means, and there is no direct behavioral evidence that the pull is being used as a representation, to indicate or stand in for something else (hence the position of the guidance theory that the bacteria are in fact too simple to represent). On the other side of the coin, as the evolutionary story shows, the pull doesn't just *happen* to lead the bacteria to oxygen-free water; there is good reason to argue that the bacteria goes down *because* that's where the oxygen-free water is. For the guidance theory, this intuition would be cashed out not directly in terms of evolutionary history, but rather in terms of an assumption, built into the bacteria, that the pull of the magnetosome carries information about the location of oxygen-free water, which assumption is a motivating reason for why the bacteria behave in accordance with the pull. As detailed in (Rosenberg and Anderson 2004; forthcoming), both "assumption of information" and "motivating reason" would be cashed out in terms of non-representational facts about how the bacteria behave in their circumstances. Just as a

[28] According to the definition of "action" given in (Rosenberg and Anderson 2004; forthcoming), the bacteria exhibit goal-directed behavior, but do not *act*. Strictly speaking, then, there are no actions to be guided by representations, and the bacteria would be ruled out by the guidance theory because they are not cognitive systems; they seem instead to be closely causally coupled to the environment, driven directly by the causal elements of the world, without the use of representations *per se*. In this respect, they are more like the slime mold slug described in (Rosenberg and Anderson, forthcoming), and appear to be another good illustration of the evolutionary pre-conditions that allowed for the emergence of representation-driven cognitive systems as a solution to the problem of behavioral guidance. Still, the example is useful in the current context as an illustration of principle.

computer is built so as to treat character strings in certain locations as if they carried information about peripheral devices like printers (and so will use the character strings that way, whatever their actual form or provenance), so the bacteria are built (by natural selection) to use the pull of the magnetosome as if it carried information about the location of oxygen-free water (and will use it this way, whatever the actual cause and direction of the pull). Likewise, just as maximizing the amount of light for photosynthesis is a motivating reason for a sunflower to track the sun, so the presence of oxygen-free water is a motivating reason for the bacteria to swim down.

That both biosemantics and the guidance theory (not to mention most of the other candidates for a theory of representation) specify the same content in most natural cases is to be expected. After all, in a properly functioning representational system, it is to be expected that a given representation will be relevantly isomorphic to, support valid inferences about, permit actions to be guided with respect to, indicate the presence of, and covary with its object; further it is likely that each of these relations will be amenable to explanation in terms of natural selection, at least in part. Where the guidance theory differs from the other candidates is in making action-guidance the fundamental determinant of content, in terms of which these other relations are to be explained. For instance, what constitutes a *relevant* isomorphism is just that isomorphism required to successfully guide the agent's action with respect to the object; likewise a given item has its evolutionary history *because* of its usefulness in guiding an agent's behavior, thereby increasing longevity and the chances of successful reproduction.

This is not to say, however, that these theories give the same answer in *every* instance. For imagine a tribe of beings with an inner structure (call it a "belief") the apparent function of which is to prevent them from eating tree frogs. Suppose further that, in their natural state, these tree frogs are poisonous, and that part of the explanation for the preservation of this "belief" is that, on average (or just often enough), beings that do not eat these frogs live longer than those who do. So far it looks like we have the beginnings of a case (by analogy with the magnetotropic bacteria) that the content of this "belief" is that tree frogs are poisonous. However, I would like to suggest that whether or not this *is* the content depends a great deal on how we fill

out the rest of the story, and that the determining factor turns out to be, not the evolutionary story, but the behavioral one; content, that is, tracks not evolutionary history but current behavioral function.

Story 1. It turns out that washing the tree frogs in salt water removes the poison. The members of the tribe have a second inner structure (another "belief") that, in conjunction with the first, causes them not to eat tree frogs until thoroughly washed in salt water. The agents carefully monitor the slipperiness of the frogs, and only eat those that do not feel very slippery. Because the ability to eat some tree frogs gives a further advantage, these structures are passed on together.

Story 2. It turns out that washing the tree frogs in salt water removes the poison. However, the members of the tribe have a second inner structure that, in conjunction with the first, causes them to listen at a certain cave when they capture a tree frog. After listening at the cave, they always release the frog. When the members of this tribe acquire the same inner structure as acquired in story 1, it does not cause them to eat washed tree frogs. Rather, they bring the washed tree frogs to the cave, whereupon they usually (but not always) release them. Because the ability to eat some tree frogs gives a further advantage, these structures are passed on together.

Story 1 seems to solidify the claim that the content of the original "belief" was indeed, something like "tree frogs are poisonous", but the claim seems to rest *not* on the evolutionary history, but rather on considerations involving the actual interaction of the inner structures in question in producing and guiding the agent's actions. For in the first case, another "belief", guiding removal of the poison, immediately allowed the members of the tribe to eat some of the tree frogs. The agents were careful to wash thoroughly, and monitored the slipperiness of the frog (which they took to be an indirect indication of the poisonousness of the frog), only eating when the slipperiness diminished sufficiently. In contrast, the second story suggests a better interpretation of the "belief" might be that "tree frogs are forbidden." These agents, rather than focusing on and monitoring the slipperiness of the frogs, instead listen at the cave to determine the will of the oracle. If the auspices were good (the frog was "clean") eating the frog was permitted, but not otherwise. What differentiates these cases is not the basic evolutionary story (we can

make this identical in each case), but the behavioral one: what the agent actually does, and what entities it monitors, in light of its "beliefs" and its circumstances.

Naturally enough, evolutionary history can often provide important insight into representational content, and will generally be a necessary ingredient in any explanation for the emergence of a particular behavior-determining mechanism, but what is not clear is whether, given the variety of ways a given fact about the world can be taken up into and given significance within a representational system, the environmental correspondences that evolution exploits are sufficient in themselves to fix representational content. The example above suggests that such correspondences will not always be sufficient, and further that where one must turn to adjudicate between different possibilities for content is not evolution, but action. And this, I should like to argue, is as it should be. It is extremely important to understanding the functions of mind that evolution in fact exploits the (relatively) stable features and relations in the world in generating and selecting various behavior-determining mechanisms. But surely identifying *which* fact or relation is being exploited is always only part of the story; what one must also determine, to fix representational content, is exactly *how*, and to what ends, a given fact or relation is being exploited. Evolutionary history can often give us the former, but is not fine-grained enough to reliably determine the latter. This latter determination must rest, not on the history of an item, but on the details of its function, and in the case of action-producing and guiding mechanisms this means coming to understand what actions are generated with what motivations, in what circumstances and with what foci.

Before bringing this essay to a close, it is important to briefly discuss one further matter. The primary basis for Millikan's dissatisfaction with the various theories of representation surveyed was their inability to properly deal with error (indeed, critiques of the sort outlined above are quite standard in the literature). Thus, if the guidance theory is to be even initially acceptable, it must provide a plausible basis for distinguishing between true and false representations. More particularly, it must be able to explain how it is possible for a representation to be in error and yet still be a (poor, false, inadequate, unfaithful) representation *of* the object to which it is apparently unfaithful.

Not surprisingly, the guidance theory explains this possibility in terms of the possibility that the actions guided by representations can fail. If a representation **R** represents entity **E** in circumstances **C** just in case **R** is standardly used by agent **A** to guide some action **B** with respect to **E** in **C**, then a false (poor, inadequate, unfaithful) representation **R** is one meeting these conditions for having a specific intentional content, however some action **B**, guided with respect to **E** by the use of **R** in **C**, will (would, if taken by **A**) fail because of **R** (Rosenberg and Anderson 2004; forthcoming).[29] Intuitively, a representation can (a) be related to a particular object in that it is standardly used to guide actions taken with respect to that object, (b) when deployed in a particular case in fact guide an action taken with respect to that object, and yet (c) nevertheless fail to be properly isomorphic[30] to the object in question under the circumstances, as required by the chosen action. In these circumstances, any action taken which relies for its success on some given isomorphism which is in fact not present, will be, to this extent, mis-guided, and will fail. For a simple, concrete example, consider my representation of my coffee pot (standardly used by me to guide my actions with respect to this coffee pot). We can imagine some particular dysfunction, perhaps in the mechanisms governing the perceptual updating of this representation, that, while not removing my general ability to take actions with respect to the coffee pot and be guided in those actions by my representation, nevertheless produce some particular inaccuracy in that representation. For instance, based on the guidance received by my representation (which, we might colloquially say, represents the pot as being open, but, more precisely, allows me to calculate the likely outcomes of possible actions, and to choose the one which leads to desired effect), I proceed to pour boiling water over that pot, with the intent of filling it. However, the pot is in fact closed, and thus my action fails in its intent, and the boiling water ends up all over my counter, instead. My representation was in error (in these circumstances for this action) with respect to the pot, but does not on this account fail to be a representation *of* the pot.

---

[29] Technically, we should say that in such a case **R** is in error with respect to **E** in circumstances **C** for action **B**. It may well be perfectly accurate with respect to **E** in some other circumstance, or for use to guide some other action. The conditions for saying that a representation is in error *simpliciter* will likely vary by circumstance (why such a claim is important or necessary, and what such a claim would be used to establish), and are not of concern to us here.

[30] Where "properly isomorphic" is defined in terms of the particular requirements of a given action in these circumstances.

One advantage of this account of error over the other possibilities on offer, including Millikan's, is that it meets Bickhard's meta-epistemological requirement that representational error be (at least theoretically, potentially) detectable by the representing system itself (Bickhard 1993, 1999).[31] As Bickhard points out, in so far as (a) representational error is cashed out in terms of deviation from proper function; and (b) proper function is understood in terms of such things as evolutionary histories; and (c) the system itself is not capable of assessing these histories so as to come to an understanding of its proper function, then (d) it does not seem that the system itself would be capable of noting when it is *mal*-functioning, nor of detecting representational error. The typical failure to provide analyses of representational error that are valid from the standpoint of the system itself is another manifestation of the third-personal perspective generally adopted by detection theories of representation, as noted above. In contrast, when error is cashed out in terms of failure of action, then so long as the system has the resources to detect when its actions fail, it likewise has the resources to detect representational error. For a detailed account of how failure of action can uncover representational error, and help guide correction, see (Anderson, *forthcoming*).

4. Conclusion

The guidance theory is an action-oriented, even, one could say, an *action-grounded* theory of representation that combines the virtues of typical internalist, externalist, and evolutionary theories of content. The guidance theory allows one to preserve the internalist insight that the functional role of an item determines its representational content, and yet is also able to provide the kind of grounding in the world for which theorists often turn to causally-based externalist accounts. It is compatible with a naturalistic, broadly functionalist, ecological view of organisms, and its basic tenets seem to follow naturally from an evolutionary approach to mind. However, while allowing an important role for evolutionary history in the

[31] As has been noted in previous work, Bickhard's theory of interactive representation is very much part of the same pragmatic, naturalistic tradition as the guidance theory, and it is motivated by the same fundamental insight regarding the epistemic importance of action and interaction. However, the guidance theory offers a significantly different development and formalization of this shared insight. For instance, Bickhard's analysis relies heavily on control theory, cashes out representational content in terms of 'environmental interactive properties', and assumes some version of process ontology. The guidance theory, while compatible with these possibilities, does not require them. In particular, it should be noted that for the guidance theory, representations are of *entities*, and the guidance theory takes no position on their ultimate ontological bases. In contrast, Bickhard's theory appears to require that entities (and ordinary objects) are in fact *defined* in terms of the 'environmental interactive properties' that compose representations; hence, one presumes, his commitment to process philosophy and it's ontological underpinnings. Still, the relative advantages of these two analyses remain largely to be determined.

fixing of representational content, that role is indirect, for it is not evolutionary history *per se* that determines content, but the function of a representation in guiding action. Since it is in virtue of their role in guiding action that the elements of an organism's cognitive systems are primarily exposed to selection pressures, this seems the proper place to locate the influence of evolutionary history on their structure and content.

References

Anderson, M. L. 2003a. Embodied cognition: A field guide. *Artificial Intelligence* 149(1): 91-130.

Anderson, M. L. 2003b. Representations, symbols, and embodiment. *Artificial Intelligence* 149(1): 151-6.

Anderson, M. L. forthcoming. Cognitive science and epistemic openness. *Phenomenology and the Cognitive Sciences*.

Bickhard, M. H. 1993. Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, 285-333.

Bickhard, M. H. 1999. Interaction and representation. *Theory & Psychology*, 9(4), 435-458.

Block, N. 1986. Advertisement for a semantics for psychology. In: P. French, T. Uehling and H. Wettstein, eds. *Midwest Studies in Philosophy* X**:** 615-678. Minneapolis: University of Minnesota Press.

Chrisley, R. 1995. Non-conceptual content and robotics: Taking embodiment seriously. In: K. Ford, C. Glymour, and P. Hayes (eds) *Android Epistemology*. Cambridge, MA: AAAI/MIT Press, pp. 141-66.

Deban, S. M., O'Reilly, J. C. and Nishikawa, K. C. 2001. The evolution of the motor control of feeding in amphibians. *Amer. Zool.* 41: 1280-1298.

Dretske, F. 1981. *Knowledge and the flow of information*. Cambridge, MA, MIT Press.

Dretske, F. 1986. Misrepresentation. In: R. Bogdan, ed. *Belief: Form, Content, and Function*. New York: Oxford University Press.

Dretske, F. 1988. *Explaining behavior.* Cambridge, MA: MIT Press.

Fodor, J. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

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

Franklin, S. 1995. *Artificial Minds.* Cambridge, MA: MIT Press.

Gibson, J.J. 1966. *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J.J. 1977. On the analysis of change in the optic array in cotemporary research in visual space and motion perception. *Scandinavian Journal of Psychology*, 18: 161-3.

Harman, G. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*, 242-56.

Harman, G. 1987. (Nonsolopsistic) conceptual role semantics. In E. LePore (Ed.), *Semantics of Natural Language*. New York, Academic Press.

Ingle, D. J. 1973. Two visual systems in the frog. *Science* 181: 1053-55.

Ingle, D. J. 1980. Some effects of pretectum lesions on the frog's detection of stationary objects. *Behavioural Brain Research* 1:139-63

Ingle, D. J. 1982. Organization of visuomotor behaviors in vertebrates. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, eds., *Analysis of visual behavior,* MIT Press.

Lettvin, J., Maturana, H., McCulloch, W., and Pitts, W. 1959. What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, 47: 1940-1951.

Lloyd, D. 1989. Simple Minds. Cambridge MA: MIT Press.

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.

Newell, A. and Simon, H. A. 1976. Computer science as empirical enquiry. *Communications of the ACM*, 19:113-126.

O'Donovan-Anderson, M. 1996. Wittgenstein and Rousseau on the context of justification. *Philosophy and Social Criticism* 22: 75-92.

Rosenberg, G. and Anderson, M.L. 2004. A brief introduction to the guidance theory of representation. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Rosenberg, G. and Anderson, M.L. forthcoming. Content and action: The guidance theory of representation.

Sandel, M. 1997. *Liberalism and the Limits of Justice*, 2ed. Cambridge: Cambridge University Press.

Schmitt, R. 1995. *Beyond Separateness: The Social Nature of Human Beings—Their Autonomy Knowledge and Power*. Boulder, CO: Wesview Press.

Taylor, C. 1989. *Sources of the Self: The Making of Modern Identity*. Cambridge, MA: Harvard University Press.

Varela, F. J., Thompson, E. and Rosch E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

White, S. 1982. Partial character and the language of thought. *Pacific Philosophical Quarterly* 63: 347-65