

## Ch 7. Animat Semantics

### 0. *Introduction: Representations for Consciousness*

There are two sorts of representations that figure in AEI—the egocentric and the allocentric—and it is the goal of this and the next chapter to flesh out accounts of these kinds of representation that are consistent with identity theory. One of the main challenges to identity theory that arises as soon as the topic of representation comes up is that representation strikes so many thinkers as requiring a relational analysis, whereby mental representations essentially involve relations between the organism that has the representations and the typically external objects and events which are the things represented. Such a relational account is at odds with the identity theory insofar as the identity theory reduces mental states and properties to wholly internal things in the organism, specifically to things in the organism’s brain.

While chapter 3’s unicorn argument did much to cast doubt on the appropriateness of relational accounts of representation for theories of consciousness, other chapters may have given the appearance that I was helping myself to a relational account. In particular, the sensational states in ch 2 with, in Churchland’s phrase, “objective intentionality”, spelled out largely in terms of the information they carry about their causes, look like relational representations of the sort dismissed in chapter 3 in the discussions of the direct reference hypothesis (DR). So, at first glance, it looks like the current theory—AEI—is infected at its lowest levels with a commitment to externalism. And if one further commits to the prima facie plausible supposition that higher levels inherit at least some of their contents from lower levels, then it looks like the externalist infection spreads to higher levels as well.

My aim is to nip this in the bud and stop the threat at its threatened point of entry. The current chapter focuses largely on low-level representations, and does so by focusing on the simplest creatures that plausibly have mental representations. It will remain for the following chapter to further address higher-level representations.

### 1. *The Economy Problem*

Representations play diverse roles in the mental lives of creatures. To mention a few: some serve as perceptions, others as memories, and yet others as intentions. It is interesting to dwell on what some of the differences consist in. Consider, for example, the difference between a memory of some perceived event and a plan or intention to do something. One mark of contrast may be made as follows. In the case of memory, the representation is of some past event, whereas in intention the representation is of some future event. Another mark of contrast may be made in terms of the different characteristic causal interactions that the representations enter into with each other as well as the rest of the organism. Intentions, being imperative representations, will thus play causal roles distinct from those played by indicative representations such as memories. Intentions will be involved in causing the organism to do something and the memory will be caused by some no past mental state. We might label these two kinds of contrast a contrast of *content* and a contrast of *causal role*, respectively.

While there are many theories of what representational content is and how representations come to have it, it is not entirely clear that these theories are compatible with basic assumptions about the diverse roles that representations play in the internal causal economies of organisms. Let us call the problem of showing the compatibility of a

theory of content and these pre-theoretic assumptions about the roles of representations within a causal economy “the economy problem”.

One way of conveying the economy problem is by criticizing the over-emphasis that perception has received in theories of content. Now, of course, focusing on perception may turn out to be a good thing and it may make especially good sense given the importance of perception in consciousness. However, perception brings with it various special problems. To get a feel for such problems, in particular, the economy problem, consider the kind of stock example typical of the content literature. Smith has a mental representation heretofore referred to as “/cow/”. As the story goes, /cow/ means cow, that is, Smith has /cow/ in his head and /cow/ represents a cow, or cow-ness, or cows in general. On the standard story, Smith will come to have a tokenning of the representation type /cow/ when Smith is in perceptual causal contact with a cow and comes to believe that there is a cow, presumably by having, in his head /there/ + /is/ + /a/ + /cow/ or some other concatenation of /cow/ with various other mental representations. The main question addressed in this literature is how /cow/, a physical sort of thing in Smith's head, comes to represent a cow, a physical sort of thing outside of the Smith's head. This focus on the perceptual case has made causal informational proposals seem rather attractive to quite a few people<sup>59</sup>, so let us focus on the following sort of suggestion, namely, that /cow/ represents cow because in typical scenarios, or ideal scenarios, or in the relevant evolutionary scenarios, /cow/s are caused by cows, that is, /cow/s carry information about cows. Thus, tokenings of /cow/s in the heads of Smith

---

<sup>59</sup> Such as Fodor (1998), Dretske (1995), Lycan (1996), and Tye (1995).

and his relatives are part of the operation of a cow-detector. A widespread presumption of this kind of view is that the /cow/s you find in the perceptual case are the same things that will be deployed in the memory, planning, and counterfactual reasoning cases too. The presumption, inherited from a long empiricist tradition, is that whatever happens in perception to wed representations to their contents can simply be passed along and retained for use in non-perceptual mental tasks. In its most literal form, this is the view that whatever happens to items in the perception “box” be sufficient to mark those items (picture them as punch cards, if you like) as bearing representational contents. Those items can thus be passed to other boxes in the cognitive economy, and retain their marks of representational content even after they may go on to play quite different causal roles. This is an interesting suggestion, but certainly open for questioning. That is, what might seem like a good idea about the nature of representations in connection with perception may not generalize to all the other sorts of things mental representations are supposed to do. Presumably, /cow/s, that is, mental representations of cows, have a lot more work to do than take part in perceptions. Consider that /cow/s are used to remember cows, to make plans concerning future encounters with cows, and to reason about counterfactual conditions concerning cows (e.g. what if a cow burst into this room right now?). Perhaps, then, the sorts of conditions that bestow representational contents onto perceptual states are very different from the conditions on representation in memory, which are yet different from the conditions for representation in planning, counterfactual reasoning, and so on.

A second concern, not unrelated to the first, is how you tell what and where the /cows/ are in the first place. Focusing on the case of perceptual belief brings with it

certain natural suggestions: point Smith at some cows and look for the brain bits that seem to “light up” the most. Much talk of representation in neuroscience is accompanied by precisely this sort of methodology. But are the bits that light up during the retrieval of memories of cows or counterfactual reasoning about cows the same bits that light up in perceptions of cows? And more to the point, how will various theories of representational content cope with the different possible answers to this question?

The economy problem might best be seen as decomposing into a pair of problems, the first concerning a question of representational content and the second concerning a question of representational vehicles. The economy problem for content is the question of whether the conditions that establish representational content for perceptual representations are the (qualitatively or numerically) same conditions that establish the representational contents of memories and intentions or whether distinct conditions are necessary. The economy problem for vehicles is the question of whether the vehicles of perceptual representations will be the (qualitatively or numerically) same vehicles as in memories and intentions or whether distinct vehicles are necessary.

## **2. *Analytic Functionalism and Apsychogenesis***

In the previous section, the difficulties of solving the economy problem were highlighted by showing how problematic it is for a theory of content like an informational theory. One thing that we might say about an informational theory that highlights why it has such a hard time with the economy problem is that it *doesn't* explain representation in terms of the roles definitive of a mental state's place in a causal cognitive economy. This diagnoses of informational theories suggests that a theory of content would have a much

easier than if it granted a constitutive role in the fixation of content to the causal relations between mental states definitive of the state's place in the cognitive economy. Thus, the central question posed by the economy problem—how can a theory of content be shown to be consistent with the diverse roles played in the cognitive economy?—receives an easy answer from any theory that simply explains content in terms of such roles, perhaps by simply *identifying* content with the causal interactions definitive of the state's place in the economy.

One such theory that will be useful to contemplate is analytic functionalism. Analytic functionalism will also be useful in contemplating the answers it gives to questions such as: What are minds? What are mental states? What makes analytical functionalists *functionalists* is their belief that what makes something a mental state is the role that it plays in a complex economy of causal interactions. What makes analytical functionalists *analytical* is their belief that which roles are essential is to be discovered by a consultation of common sense knowledge about mental states. Thus, as Braddon-Mitchell and Jackson (2000) write:

We distinguish the roles that matter for having a mind, and matter for being in one or another mental state, by drawing on what is common knowledge about mental states. We extract the crucial functional roles from the huge collection of what is pretty much common knowledge about pains, itches, beliefs, desires, intentions, and so on and so forth, focusing on what is most central to our conception of what a pain is, what a belief is, what it is to desire beer and all the rest. And we can group the common knowledge into the three clauses distinctive of the functionalist approach. The input clauses will contain sentences like 'Bodily

damage causes pain' and 'Chairs in front of people in daylight cause perceptions as of chairs'; the output clauses will contain sentences like 'Pain causes bodily movement that relieves the pain and minimizes damage' and 'Desire for beer causes behaviour that leads to beer consumption'; the internal clauses will contain sentences like 'Perception as of beer in front of one typically causes belief in beer in front of one' and 'Belief that if p then q typically causes belief that q on learning p'.

...

What then is a given mental state M, according to the common sense functionalist story? It is the state that plays the M role in the network of interconnections delivered by common knowledge about the mind. The network in effect identifies a role for each mental states, and thus, according to it, a mental state is simply the state that occupies the role definitive of it (pp 45-47).

Whatever attractions analytic functionalism might hold as a solution to the economy problem, as well as a source of an internalistic theory of content, it is not without problems of its own. In particular are the following three serious and related problems.

The first problem is that analytical functionalism appears to be committed to the existence of analytical truths and various philosophers inspired by Quine have been skeptical of analytical truths. As Prinz (2006) succinctly sums up this Quinean skepticism, the objection is that "[r]oughly, definitions of analyticity are either circular because they invoke semantic concepts that presuppose analyticity, or they are epistemically implausible, because they presuppose a notion of unrevisability that cannot

be reconciled with the holistic nature of confirmation" (p. 92). There are two main ways in which analytic functionalism seems saddled with belief in analytic truths. The first is concerns the nature of psychological state types such as beliefs and desires. Analytical functionalism is committed to there being analytic truths concerning the necessary and sufficient conditions for being a belief. The second concerns the meaning of mental representations. The most natural theory of meaning for the analytic functionalist to adopt is that what makes one's belief about, say, cows, have the meaning that it does, is the causal relations it bears to all other belief states. However, it is likely that no two people have all the same beliefs about cows. Thus, on pain of asserting that no one means the same thing when they think about cows, the functionalist cannot allow that every belief one has about cows affects the meaning of one's cow thoughts. In order to allow that people with divergent beliefs about cows can both share the concept of cows, that is, both think about the same things when they think about cows, the analytic functionalist seems forced to draw a distinction between analytic and synthetic beliefs, eg., a distinction between beliefs about cows that are constitutive of cow concepts and beliefs that are not. But if Quinean skepticism about the analytic/synthetic distinction is correct, no such distinction is forthcoming.

The second problem arises from worries about how minds are implemented in brains. Many so-called connectionists may be seen to agree with analytical functionalists that mental states are defined in terms of networks. However, many connectionists may object that when one looks to neural network implementations of cognitive functions, it is not clear that the sets of nodes and relations postulated by common sense psychology will map on to the nodes and relations postulated by a connectionist architecture (see, e.g.

Ramsey, et al., 1991). The question arises of whether folk-psychological states will smoothly reduce to brain states or be eliminated in favor of them. (I will not discuss further the third option that folk psychological states concern a domain autonomous from brainstates.) Related are the sorts of concerns about functionalism raised in chapter 1.

A third problem arises from worries about the evolution of cognition. If a mind just is whatever the collection of folk psychological platitudes are true of, then there seem not to be any simple minds, for a so called simple mind would be something that the folk psychological platitudes were only partially true of in the sense that only some proper subset of the platitudes were true of it. However a very plausible proposal for how our minds evolved is from simpler minds. It counts against a theory that it rules out a priori the existence of simpler minds than ours for it leaves utterly mysterious what the evolutionary forebears of our minds were.

Problem 2 receives a direct solution by adopting a connectionist theory of mind. Problems 1 and 3 receive solutions by utilizing a similarity metric between minds made possible by adopting connections. Connectionism was described earlier in ch xx. The similarity measure is as follows.

Laakso and Cottrell (1999, 2000) propose a method whereby representations in distinct networks may be quantified with respect to their similarity. Such a similarity measure may apply even in cases where the networks in question differ with respect to their numbers of hidden units and thus the number of dimensions of their respective vector spaces. In brief the technique involves first assessing the distances between various vectors within a single network and second measuring correlations between

relative distances between points in one network and points in another. Points in distinct networks are highly similar if their distinct relative distances are highly correlated.

As Churchland *zx* has pointed out regarding the analytic/synthetic distinction related worries, the Laakso and Cottrell technique allows one to bypass attributions of literally identical representations to distinct individuals and make do instead with objective measures of degrees of similarity between the representations of different individuals. Thus if I believe that a cow once ate my bother's hat and you have no such belief, we may nonetheless have measurably similar cow concepts. This is no less true of our psychological concepts such as our concepts of belief and concepts of desire. The so-called common-sense platitudes of folk psychology so important to analytic functionalism may very well diverge from folk to folk and the best we can say is that each person's divergent beliefs about beliefs may be similar. And similarity measures are not restricted to the concepts that constitute various folk theories, we may additionally make meaningful comparisons between various folk theories and various scientific theories. This last maneuver allows us to retain one of the key insights of analytic functionalism mentioned earlier: that we are in need of some kind of answer to the question 'how do you know that your theory is a theory of belief?' The answer will be along the lines of "because what I'm talking about is similar to beliefs."

Regarding the question of simple minds, if there are no analytic truths, then there is no *a priori* basis (if any) for drawing a crisp boundary between the systems that are genuine minds and those that are not. Similarity measurements between simple minds and human minds would form the basis for a (mind-body?) continuum along which to position various natural and artificial instances. How useful for understanding human

minds will be the study of systems far away on the continuum? We cannot know *a priori* the answer to such a question.

Of course, one might wonder why one should even care about such a question. My motive for caring about the evolutionary background of the simplest cognitive behaviors emerges from the following presumptions. I presume, and am unlikely alone in doing so, that the simplest forms of intelligent behaviors are adaptive. That is, intelligent behaviors, at least of the simplest varieties, provide biological benefits to the organisms that perform them. I presume also that just as there was a time in the history of the universe that there were no biological organisms, there was a time in the history of the universe that there were no organisms performing intelligent behaviors. Since “abiogenesis” is the term referring to the hypothesized emergence of life from non-living matter, we coin the term “apsychogenesis” to refer to the hypothesized emergence of intelligence from non-intelligent systems. When, in the history of the universe did abiogenesis and apsychogenesis occur? No one knows, but we doubt that apsychogenesis preceded abiogenesis. They either coincided or abiogenesis occurred first. However, the latter option strikes us as the more plausible of the two. Adding to our growing list of presumptions, we further presume that the problem of understanding apsychogenesis is best understood in the context of an evolutionary framework. Thus we are led to ask: What pressures applied to non-intelligent organisms yielded the earliest and simplest forms of intelligence? If mental representations are to underwrite intelligent behavior, then questions of the evolvability of intelligence will be closely related to questions of the evolvability of mental representations.

I will tackle the topics of intelligence, representation, and evolution by examining computer simulations of evolved adaptive behaviors. The simulated organisms, behaviors, and environments will be simple enough to make tractable questions concerning the physical properties that constitute representations and the roles representations play in adaptive intelligent behaviors.

### 3. *Animat Methodology*

The strategy of examining the simplest examples of intelligent behavior by examining the simplest examples of organisms that behave intelligently confers the following advantage. The simpler the creature the easier it will be to keep track of the creature's internal structure, the structure of the creature's environment, and any relations between the two kinds of structure that might be crucial to an account of how representation is effected and intelligence is enabled. Additionally we will be in a better position to assess whether there are plausible candidates for relations between internal and external structures in virtue of which the former count as representations of the latter. Further, dealing with extremely simple cases will allow for tractable computer simulations of creature behavior as well as simulations of the evolutionary forces that contribute to the emergence of such behaviors.

The above questions concerning representation are pursued here by employing a cognitive scientific methodology come to be known recently as bottom-up AI or the animat approach (for a review see Guillot and Meyer 2001). An animat is an artificial animal, either computer simulated or robotic. As I shall define this methodology, animat methodology involves three characteristic explanatory strategies: synthesis, holism, and

incrementalism. The synthetic element involves explaining target phenomena by attempting to synthesize artificial versions of them, a characteristic inherited in large part from earlier versions of Artificial Intelligence (Good Old Fashioned Artificial Intelligence (GOFAI) as well as connectionist approaches). The holism referred to here is not necessarily restricted to the semantic holism familiar in other areas of philosophy of mind or cognitive science<sup>60</sup> but is instead concerned with function more generally. The holistic take on function is that the function of an organ or a behavior is best understood in the context of the whole organism, or, more broadly still, in the context of the organism's physical and/or social environment. It is thus both embodied and embedded (Clark 1997). However, this holistic impulse might seem to conflict with attempts to synthesize phenomena. Synthesis must simplify to be tractable, yet whole organisms are more complex than their subsystems, and social systems and ecosystems are even more complex. An older strategy of simplification involves focusing on subsystems of human cognitive processes, for example, as was done in GOFAI and connectionist models of word recognition. The comparatively newer strategy of simplification embraced by the Animat approach involves focusing on the entirety of organisms much simpler than the human case, thus heeding Dennett's rallying cry/question, "Why not the whole iguana?" (1998: 309). In animat research, projects of synthesis involve modeling the simplest intelligent behaviors such as obstacle avoidance and food finding by chemotaxis. The incrementalism of the animat approach involves building up from these simplest cases to the more complex via a gradual addition of complicating factors, as in, for instance,

---

<sup>60</sup> As discussed, for instance, in Fodor and Lepore (1992).

roboticist Rodney Brooks' (1999) ongoing project of building an incrementalist bridge from robotic insects like Attila through to the humanoid robot, Cog.

Some of the earliest practitioners of animat methodology did not emanate from the engineering and computer sciences, but were instead neuroscientists. The neuroscientists Grey Walter (1963) and Valentino Braitenberg (1984) have had a deep impact on the practice of animat methodology. Walter built his robotic "turtles" Elmer and Elsie out of vacuum tubes and other electric components of the day. Elmer and Elsie were wheeled animats with perceptual sensitivity to light and sound and capable of a rudimentary form of associative learning. Unlike Walter, Braitenberg did not implement his ideas in hardware, but the thought experiments conducted in *Vehicles: Experiments in Synthetic Psychology* inspired the projects of many robotocists. Braitenberg's animats, the vehicles of his book's title, were envisioned as relatively simple collections of sensors and motors with excitatory and inhibitory connections between them. Figure 1 depicts two of Braitenberg's simplest vehicles in the proximity of a stimulus.

<<INSERT FIGURE 1 HERE >>

**Figure 1. Two Braitenberg vehicles and a stimulus source. Figure drawn by Pete**

**Mandik.**

The vehicle on the left has a single sensor on its front connected to a single motor in its rear. If the line connecting the sensor to the motor is excitatory, then increased sensor activity will result in increased motor activity. Stimulation of the sensor will result in the vehicle accelerating toward the stimulus. The vehicle on the right has two sensors with

crossed connections to two motors. If the connections are excitatory, the vehicle will turn toward a stimulus. For example, if the stimulus is to the right of the vehicle this will result in higher activity in the right sensor than the left sensor, resulting in higher activity in the left motor than the right motor. If, in contrast, the excitatory connections are parallel and not crossed, then the creature will move away from the light.

Two points of immediate concern, both discussed by Braitenberg himself, fall out of a consideration of Braitenberg's vehicles. The first is the question of to what degree it seems natural to attribute psychological states to the vehicles, for example, to describe these creatures as loving or fearing the stimulus source.. The second is the question of whether relatively simple systems can give rise to models of coherent behaviors such as taxis (the movement toward or away from a stimulus source) and kinesis (movement triggered by a stimulus)—that is, the question of whether these systems are amenable to neuroscientific and neuroethological description. The terms “sensor”, “motor”, “excitatory connection” and “inhibitory connection” have natural applications in the neurosciences. And the promise of seeing how they work together in the context of an entire organism to give rise to survival promoting behaviors like food finding by positive phototaxis or chemotaxis sparks the hope that along this path lies accounts of the evolutionary function of the earliest brains and nervous systems more generally.

Contemporary practitioners of animat methodology have at their hands techniques for addressing these evolutionary questions that arise. Nowadays many computer programs exist that allow for the evolution of minimally cognitive behaviors in populations of relatively simple neural network controlled critters (for a review see Taylor and Massey 2001). Such programs allow for the simulation of evolution by natural

selection by providing for the mechanisms of the variable inheritance of fitness. Such programs allow for simulations that capture the embodied, embedded, and evolutionary aspects of cognition.

The point of the simulations described below is to show that relatively simple autonomous agents—agents with neural controllers of only, for example, a dozen neurons and neural connections—are capable of acquiring and sustaining in an evolutionary context several varieties of mental representation. The successes of these simulations have implications for addressing the economy problem.

#### **4. *Representation-hungry Problems for Food-hungry Worms***

##### **4.1 When You Gotta Represent**

Let us take a brief look at a folk-psychological explanation of a piece of intelligent behavior. Consider George. George is opening a refrigerator. Why? What explanation is available for this action? A folk-psychological explanation will advert to a collection of psychological states that jointly constitute a cause of George's behavior. An example collection of such states would include a desire, a perception, and a memory. One explanation of Georges' behavior then would advert to George's *desire* to drink some beer, George's visual *perception* that there is a refrigerator in front of him, and George's memory that he put some beer in the refrigerator the day before.

There are a few useful points to note about this explanation. First, the psychological states are not individually sufficient to cause a behavior, but must act in concert. A belief that there is beer in front of you will contribute to causing you to move toward it if combined with a desire for beer and will contribute to causing you to move

away from it if combined with a fear of beer. Similarly, a desire for beer will contribute to causing you to move forward if combined with a belief that beer lies ahead and cause you to move in some other direction if combined with some other belief. In summary, psychological states contribute to the causes of behavior by acting in concert.

A second useful point to note about this sort of explanation is that the psychological states are identified in part by their representational content and in part by what attitude the person is taking toward that content. In the case of George's memory that he put some beer in the refrigerator, the representational content of the memory is that George put some beer in the refrigerator and the attitude is one of remembering. Different types of attitude can be taken toward one and the same content (e.g. *remembering* buying beer; *planning on* buying beer) and one and the same attitude type can be taken toward different contents (e.g. perceiving *that there is a beer in front of me*, perceiving *that there is a slice of pizza in front of me*). In summary, psychological states that are causes of intelligent behaviors admit of a distinction between their representational contents and the attitude that is taken toward those representational contents.

A third useful point to note about these sorts of explanation is that we can make attributions of content without explicit knowledge of what, in general, representational content is. We construct such explanations on the fly without knowing, for example, what the right theory of content is or even having a theory of content (explicitly or consciously) in mind. We plan to exploit this in what follows. We will present relatively clear cases of synthetic organisms that behave in ways explainable in terms of representational states and we will do so before offering a definition of what

representations are or what representational content is. This leaves open to empirical investigation what the best accounts of representation and content are as opposed to a matter that must be settled a priori before such investigations take place.

## 4.2 The Diet of Worms

Chemotaxis—directed movement in response to a chemical stimulus—is one of the simplest forms of organism behavior. It is an adaptive behavior as when, for example, positive chemotaxis is used to move toward a food source or negative chemotaxis is used to move away from a toxin. The underlying mechanisms chemotaxis are relatively well understood and amenable to modeling and simulation (Mandik 2002, 2003, 2005)

Chemotaxis is appropriate to regard as cognitive. As we will argue below, it constitutes what Clark and Toribio (1994) call a “representation hungry” problem. To appreciate the informational demands that chemotaxis places upon an organism, it is useful to consider the problem in the abstract. The central problem that must be solved in chemotaxis is the navigation of a stimulus gradient, and the most abstract characterization would be the same for other taxes such as thermotaxis or phototaxis. To focus on a simplified abstract case of positive phototaxis, imagine a creature traversing a plane and utilizing a pair of light sensors—one on the left and one on the right. Activity in each sensor is a function of how much light is falling on it in such a way that the sensor closer to the source of light will have a greater degree of activation. Thus, the difference in the activity between the two sensors encodes the location of the light source in a two dimensional egocentric space. Information encoded by the sensors can be relayed to and decoded by motor systems responsible for steering the creature. For example, left and right opposing

muscles might have their activity be directly modulated by contralateral sensors so that the greater contraction corresponds to the side with the greatest sensor activity, thus steering the creature toward the light.

Consider now the problem of phototaxis as confronted by a creature with only a single sensor. The one-sensor creature will not be in a position to directly perceive the direction of the light since activity in a single sensor does not differentiate from, say, light being three feet to the left or three feet to the right. Of course, the creature might try to exploit the fact that the sensor is moving and make note of changes in sensor activity over time, but such a strategy will be available only to creatures that have some form of memory. Exploiting the change of sensor activity will require a means of comparing the current sensor activity to some past sensor activity.

Note the folk-psychological explanation of how a human would solve the problem of one-sensor taxis. To imagine that you are in a gradient it will do to imagine that you are literally in a fog so dense that while you can ascertain how dense it is where you are, you cannot ascertain in which direction the fog gets more dense and in which direction it gets less dense. However, as you start walking after a point you notice that the fog is much less dense than it was previously. By comparing your current perception of a less dense fog to your memory of a more dense fog against the background of your knowledge that you have been walking, it is reasonable for you to infer that you are moving out of the area of greatest concentration. Conversely, if your current perception reveals a greater concentration of fog than remembered, it is reasonable for you to infer that you should turn around if you want to get out of the fog.

There are several points we should get from the above discussion. The first is that the informational demands of one-sensor chemotaxis can be readily appreciated from the point of view of folk-psychological explanation. The same point of view allows us to construct possible solutions to the problem of one-sensor chemotaxis: a creature that is both able to perceive the current concentration and remember the past concentration is thus in the position to make an inference about whether to keep moving ahead or turn in order to reach a desired location in the gradient.

One-sensor chemotaxis is accomplished by natural organisms. One particularly well studied example is the nematode worm *Caenorhabditis Elegans* (*C. Elegans*). Despite having four chemosensors, a pair in the head and a pair in the tail, there are good reasons to believe that the worm effects one-sensor, not four-sensor, chemotaxis (Feree and Lockery 1999). First off, worms are able to effect chemotaxis even when their tail sensors are removed. Second, two sensors in the head are too close together for there to be an appreciable difference between the activity in each of them in response to local concentration of attractant. Third, when navigating chemical gradients on the two dimensional surface of a Petri dish, the worms are positioned on their sides with the pair of head sensors orthogonal to the gradient. Fourth, artificial neural network controllers inspired by the neurophysiology of *C. Elegans* with only a single sensor input are able to approximate real chemotaxis behaviors in simulated worms. These simulations are especially interesting to examine in some detail.

We next briefly review work done in simulating *C. Elegans* chemotaxis in Shawn Lockery's lab at the University of Oregon Institute of Neuroscience. In particular we focus here on work reported in Feree and Lockery 1999 and Dunn et al 2004. Ferreé and

Lockery construct a mathematical model of the control of *C. Elegans* whereby the time-derivative of the chemical concentration is computed and used to modulate the turning rate of the worm in the gradient. One of our purposes in reviewing this work is to point out how it, at best, supplies only a partial explanation of how the actual nervous systems of *C. Elegans* regulates chemotaxis. Ferree and Lockery begin by constructing a model network that makes many simplifying assumptions about the neuroanatomy and neurophysiology of the relevant circuits in *C. Elegans*. They hypothesize that the worm must “assess the gradient by computing the temporal derivative of concentration as it moves through the chemical environment” and that the behavioral upshot of this assessment is that the worm “attempts to keep its head pointed up the gradient”. Their model network consists of five neurons whose various states of activation model voltage. The single sensory input has a state of activation that reflects the local concentration of the chemical attractant. Two output neurons model the voltages of dorsal and ventral motor neurons whose relative voltages determine the worm’s neck angle. The remaining three neurons are interneurons. Each of the five neurons is connected to every other neuron by both feed-forward and feedback connections thus making a recurrent network. Ferree and Lockery optimized network parameters by using a simple simulated-annealing training algorithm to maximize a fitness function defined in terms of the change of chemical concentration. The optimized network resulted in simulated worm behavior similar to that of real worms: “oriented movement up the gradient and persistent dwelling at the peak” However, Ferree and Lockery point out that it is not obvious how the networks are effecting these behaviors: “Simple inspection of the parameters...does not necessarily lead to an intuitive understanding of how the network functions, however,

because the neural architecture and optimization procedure often favor a distributed representation of the control algorithm.” To derive “an intuitive mathematical expression for this algorithm” they manipulated the analytic solution to the linear system of equations that comprise their mathematical model. The analytic solution for the linear recurrent network is an equation wherein the rate turning is equal to the sum of a turning bias and the cumulative effect of chemosensory input on the rate of turning. This equation produces exactly the same response to chemosensory input as the original optimized network. In order to “further improve our intuition about chemotaxis control in this model” Ferree and Lockery produce a Taylor expansion of the equation in time-derivatives of the input. The extracted rule for chemotaxis control equates rate of turning with a sum whose first term is a turning bias, the second term is the zeroth time derivative of chemical concentration, the third term is the first order time derivative of chemical concentration, the fourth term is the second-order time derivative of chemical concentration, and so on. Next they compared simulated behavior wherein only some of the terms are kept. With just the turning bias and the zeroth order term, the resultant behavior was not chemotaxis but instead just a circular motion around the starting position. Adding the first order term resulted in chemotaxis as did adding the first and second order terms. Likewise did adding the first order but omitting the zeroth-order term.

Ferre and Lockery describe their accomplishment as follows: “Using analytical techniques from linear systems theory, we extracted computational rules that describe how these linear networks control chemotaxis”. However, we find the resultant mathematical *descriptions* unsatisfying insofar as they do not constitute *explanations* of

how the networks effect chemotaxis. And they do not constitute explanations because too little has yet been said about what the underlying mechanisms are and how it is that they are functioning. When we say that they do not supply a complete a complete account of the mechanism by “mechanism” we intend it in the sense of... e.g. Craver 2001 :

“Mechanisms are collections of entities and activities organized in the production of regular changes from start or set up conditions to finish or termination conditions.” p.58.

See also Craver and Darden 2001, Machamer, Darden, Craver 2000, and Bechtel and Richardson 1993)

To get a feel for what we think is still missing, recall the earlier discussion between the difference between two-sensor chemotaxis and one sensor chemotaxis. In the case of two-sensor chemotaxis, the difference in activity between the left and right sensors can be straightforwardly exploited by a steering mechanism that would guide the animal right up the gradient. For example, left and right steering muscles could be connected to the sensors in such a way that the greater activity in the right sensor will result in a greater contraction in the right steering muscle thus turning the head of the worm toward the direction of the greatest concentration. If the worm’s head is pointed directly in the direction of the greatest concentration then the activity in the left and right sensors will be approximately equal as will be the amount of contraction in the left and right steering muscles, thus keeping the worm on course. In this description of the two-sensor case, we have at least a sketch of what the mechanism are underlying chemotaxis is. We are not in a comparable position yet with Ferree and Lockery’s mathematical description. The computation rule tells us *that* the time derivative of the concentration is being computed, but we are not yet in a position to see *how* it is being computed. We

know enough about the underlying mechanisms to know that there is sufficient information present upon which to compute the time derivative, because we know that the chemical concentration detected by the sensor is changing over time as the worm moves through the environment. However, we need to know more than that the information is *there*. We need to know how the information is *encoded* and subsequently *used* by the organism. As Akins (2001) puts a similar point:

Information that is carried by, but not encoded in, a signal is information that is available only *in theory*. To say that the information is *present* is to say only that there exists a computable function which, if used, would yield the correct result...It is present, as it were, from the point of view of the universe. But no creature has ever acted upon information that is available only in principle. (p. 381)

Lockery and his colleagues are not blind to this sort of shortcoming. In a subsequent publication Dunn et al 2004 write “The chemosensory neurons responsible for the input representation are known...as are the premotor interneurons for turning behavior...Much less is known about the interneurons that link chemosensory input to behavioral output.” To get a further handle on what the interneurons might be doing, Dunn et al run simulations of networks optimized for chemotaxis. The networks in these simulations have a single input neuron, one output neuron and eight interneurons. All of the neurons in each network are connected to each other and have self-connections as well. After optimization and testing the networks that performed successful chemotaxis were subjected to a pruning procedure whereby unused neurons and connections are eliminated. Dunn et al report that the pruned yet still-successful networks have only one

or two interneurons and they all have inhibitory feedback among all of the neurons. Dunn et al proposed that the main function of this feedback is “to regulate the latency between sensory input and behavior” but we note that while this latency regulation may indeed be occurring, it certainly does not explain how successful chemotaxis is accomplished. The mere introduction of a delay between input and response surely cannot suffice for successful chemotaxis. We hypothesize that the crucial yet underappreciated mechanism in the successful networks is the existence of recurrent connections. Recurrence has been noted by many authors (Mandik 2002, Churchland 2002, Lloyd 2003) as a mechanism whereby a network may instantiate a form of short-term or working memory, since activity in the network will not simply reflect the information currently coming into the sensory inputs, but also reflect information feeding back and thus representing past information that came into the sensory inputs. We hypothesize, then, that the recurrence is implementing a form of memory that allows the network to compute the time derivative of the concentration in virtue of both encoding information about the current concentration (in the state of the sensor) and encoding information about past concentration (in the signal propagated along the recurrent connections).

To test this hypothesis, we conducted our own simulations of *C. Elegans*' single-sensor chemotaxis. For our simulations we utilized the Framsticks 3-D Artificial Life software (Komosinski 2000) that allowed for the construction and testing of worms in a simulated physics and the optimization of the worms using simulated Darwinian selection. The morphologies of our synthetic worms are depicted in figure 1 and their neural network topologies are depicted in figure 2. Networks are modular. One module constitutes a central pattern generator that regulates forward motion by sending a

sinusoidal signal to the chain of muscles that control flagellation. Another module regulates steering with a single sensory neuron, three interneurons and one output neurons. This five neuron steering network is recurrent with every neuron in it connected to every other.

[[INSERT FIGURE 1 ABOUT HERE]]

Figure 1. Synthetic C. Elegans. On the left, front view. On the right, top view.

[[INSERT FIGURE 2 ABOUT HERE]]

Figure 2. Neural network for the synthetic C. Elegans. Neurons include one sensor (s) and several motor neurons (m) and interneurons (i). Single-headed arrows indicate flow of information from one neuron to the next. A double-headed arrows between two neurons indicates both a feed forward and a feed back connection between them.

In our simulations the initial morphologies and network topologies were set by hand. The connection weights are optimized through a Darwinian process whereby mutations are allowed only for connection weights and not to morphologies or network topologies.

Fitness is defined in terms of overall life time distance for this forced the worms both to maintain a high velocity and also to extend their lives by replenishing their energy store with found food. We compared the performance of 3 kinds of orientation networks: fully recurrent networks with sensory input, recurrent networks with no sensory input (“blind”

networks), and strictly feed-forward networks with sensory input. Four populations of each of the 3 kinds of orientation networks were subjected to the evolutionary simulation for 240 million steps of the running program.

Results are shown in figure 3 of the lifetime distances averaged over the four populations for each of the three kinds of orientation networks. The performance of the blind networks the maximal distance accomplished by worms with maximally optimized velocities but no extension of lifespan through finding food beyond whatever food they collided with accidentally. Worms with sensory inputs and recurrent connections were able to maximize their lifespan through food-finding by chemotaxis. Further, their swimming behaviors were similar to those exhibited by real *C. Elegans*: directed movement up the gradient and dwelling at the peak. Worms without recurrent connections are conferred no advantage by sensory input. Our explanation of this is that without the recurrent connections to constitute a memory, the worms are missing a crucial representation for the computation of the change of the local concentration over time. We turn now to examine the nature of these underlying representations

[[INSERT FIGURE 3 ABOUT HERE]]

Figure 3. Results of the experiment comparing recurrent, feed-forward, and blind networks in an evolutionary simulation of chemotaxis.

## 5. *Simple but Non-atomic: Representation in Structure*

We admittedly do not yet have a complete explanation of *C Elegans* chemotaxis, but we do have a pretty good sketch of what is going on: Heading in the gradient is determined by a computation that takes as inputs both a sensory representation that encodes information about the current local concentration and a memory representation that encodes information about the past local concentration. The existence of a memory mechanism was predicted by the folk psychological explanation of how one sensory chemotaxis could be effected and supported by experiments with simulated networks sufficient for chemotaxis. Further, we are in a position with respect to these models to make some remarks about what the representations are and what determines the representational contents. In the orientation networks we may discern three types of representations: sensory representations, memory representations, and motor representations. The sensory representations are states of activations in the chemo-sensory input neuron. The memory representations are signals conveyed along recurrent connections. Motor representations are states of activation in motor neurons. The contents of the representations are the things that are represented. In the sensory case, what is represented is current local concentration. In the memory case, what is represented is past local concentration. In the motor case, the representation is a command signal and what is represented is a level of muscular contraction

It is tempting at this point to construe representation as a relation. The question then arises of what the relation is between representation and represented is such that the former is a representation of the latter. Two major sorts of suggestion common in the philosophical literature on representational content seem initially applicable to the case of

the chemotaxis networks: informational approaches and isomorphism based approaches. The first sort of suggestion is that the relations that underwrite representation are causal-informational. On such a suggestion, it is in virtue of being causally correlated with a particular external state that a particular internal state comes to represent it. In the chemotaxis examples, there are indeed relations of causal correlation between the representations and what they represent. In the case of the sensory representation, there is a reliable causal correlation between the sensor state and the current local concentration and in the memory case there is a reliable causal correlation between the recurrent signal and the past local chemical concentration. The informational view must give a slightly different treatment of motor representations since commands are the casual antecedents of their representational targets.

The isomorphism suggestion seems applicable as well, though before discussing its application we need to spell out the relevant notion of isomorphism. An isomorphism is a structure preserving one-to-one mapping. A structure is a set of elements plus a set of relations defined over those elements. So, for example, a set of temperatures plus the hotter-than relation constitutes a structure as does a set of heights of a mercury column in a thermometer and the taller-than relation. A one-to-one mapping between temperatures and height just in case for any height and the next higher one they are mapped respectively to a temperature and the next hottest one.

Information-based theories of representational content make it a necessary condition on a representation  $r$  of a thing  $c$  that  $r$  carry information about (causally correlate with)  $c$ . Isomorphism-based theories of representational content make it a

necessary condition on a representation  $r$  of a thing  $c$  that  $r$  and  $c$  be elements in structures wherein an isomorphism obtains that maps  $r$  to  $c$ .

Can we adjudicate between the informational and isomorphism suggestions? More specifically, can the way in which attributions of representation in the explanations of the network control of chemotaxis be used to favor information-based theories over isomorphism-based theories or vice versa? We see the respective roles of the notions of representation, information, and isomorphism in this context as follows. The sensory and memory states are able drive *successful* chemotaxis in virtue of the informational relationships that they enter into with current and past levels of local chemical concentration, but they are able to enter into those informational relations because of their participation in isomorphisms between structures defined by ensembles of neural states and structures defined by ensembles of environmental states. In brief, in order to have representational contents that they have they must carry the information that they do and in order to carry the information that they do they must enter into the isomorphisms that they do. To spell this out a bit further will require spelling out two things: First, why it is that representation requires information and second, why information requires isomorphism.

We begin with the reason why representation requires information. A large part of the reason representation requires information in the example of the chemotaxis networks is because of the sorts of representation that we are talking about, namely sensory and memory representations. It is part of the nature of sensory states that they carry information about the current local situation of an organism and part of the nature of memory states that they carry information about the past. Another way to appreciate the

carrying of information is to realize that if the networks didn't encode information about the current and past chemical concentrations then they would not be able to give rise to the successful chemotaxis behavior. Consider the blind worms: they were deprived of the means of encoding information about the present chemical concentration. Consider also the worms with strictly feed-forward networks. Without recurrent connections, they were deprived of the means of encoding the relevant information about the past. It seems that the crucial aspect of attributing sensory and memory representations in explaining successful one-sensor chemotaxis is that such attributions track the information-bearing properties of the states.

To see why isomorphism is important, it helps to begin by considering how hard it would be to not have isomorphism. First off, note that, as Gallistel (1990) has pointed out, a one-to-one mapping can be considered as structure preserving even if the structures involved are defined in terms of sets of elements and only the identity relation. On such schemes the resultant representations are what Gallistel calls "nominal representations." For example, the set of numbers assigned to players on a sports team a set of nominal representations in this sense. There is a one-to-one mapping between numbers and players and the only relation between numbers that is mapped onto a relation between players is identity: one and the same number can only be mapped onto one and the same player. Larger numbers, however, need not indicate larger players, or heavier player and so forth.

Setting aside identity-based nominal representations as genuine isomorphisms, there is still a serious difficulty the informational theorist faces concerning the alleged dispensability of isomorphism. Even if there were a logically possible scheme that had

information without isomorphism, it is incredibly difficult, if not impossible, for such a scheme to be evolved or learned. We can see the point concerning evolution in the context of the synthetic *C. Elegans* in our artificial life simulations. Organisms' bodies, as well as the environments they are situated in, contain many physical systems that have states that fall into natural ordering relations. Consider, for example, that chemical solutions can be more or less concentrated, or that neural firings can have higher or lower rates or higher or lower voltages. It is hard, if not impossible, to see how there could be a counter-example to the following claim: Any situation in which a particular level of neural activation can be used to carry information about a particular level of chemical concentration is also going to be a situation in which a higher level of neural activation can be used to carry information about a higher level of chemical concentration. In other words, organisms and their environments are rich in structures and it is hard to see how elements in those structures can be evolved to enter into informational relationships without the structures themselves also entering into isomorphism relationships.

While our argument is, to our knowledge, unique, it is worth mentioning certain similarities between our argument, which is specifically about evolution and some other arguments that focus on learning that have appeared in the literature on isomorphism. Cummins 1996 and Churchland 2001 both endorse isomorphism based theories of representational content and both argue that a creature can only be in a position to have states that carry information about external states if the creature's internal states are embedded in a network of internal states that may be regarded as constituting knowledge of or a theory of the target domain. As Cummins (1997) puts the point:

Distal properties generally cannot be directly transduced. Instead, the detection of distal properties must be mediated by what we might as well call a theory about that property. To detect cats (an instantiation of catness) requires a theory that says, in effect, what sorts of proximal stimuli are reliable indicators of catness. To detect cats visually, you have to know how cats look. The same goes for colors, shapes, and sizes: for these to be reliably detected visually under changes in perspective, lighting, and distance requires knowledge of such facts as that retinal image size varies as the inverse square of the distance to the object. Much of the knowledge that mediates the detection of distal properties must be acquired: we are, perhaps, born with a tacit knowledge of Emmert's Law, but we are not born knowing how cats look, or with the ability to distinguish edges from shadows. (p.536-537) <sup>61</sup>

One sort of objection that we've encountered in various personal communications is that the notion of isomorphism employed above should instead be replaced with the notion of homomorphism where, in brief, the main difference between the two is that where isomorphisms involve one-to-one mappings, homomorphisms involve mapping one structure into (not onto) another. Homomorphism comes up in the literature on non-

---

<sup>61</sup> For an argument similar to Cummins' see also Churchland 2001, pp 131-132. For an argument that a creature can extract information from a perceptual representation only if certain isomorphisms obtain between states of perception and what is perceived, see Kulvicki (2004). See Sellars (xx) about how we couldn't have knowledge of the external world if not for isomorphisms between impressions and perceptible properties.

mental representations such as scientific representation (Suarez 2003) and the representations pertinent to measurement theory (Matthews 1994, Suppes and Zinnes 1965) but we think that isomorphism is more appropriate for mental representation. Trying to utilize the notion of homomorphism for mental representation would involve the idea that structure A represents B if and only if B is homomorphic to A which involves B mapping into (not onto) A. This allegedly allows for, among other things, A to be a partially accurate model of B. We might think of these mappings as, for example a mapping of physical objects or an empirical system into the real numbers allowing us to say that numbers represent physical objects.

One problem with the above homomorphism based suggestion is that we don't simply want to establish a relation between two sets: the representations and the represented. We want instead to establish a set of relations, more specifically, a set of relation that will allow us to say, for example, of each height of the mercury column, whether it represents a temperature and, if so, which one and of each temperature, if it is represented by a height of the mercury column and, if so, which one. We especially want to avoid attributing multiple contents to one and the same representation, as in, saying of a height of the mercury column that it represents multiple temperatures.

Attributing representations to an organism must involve partitioning the state-space of the organism and the state-space of its environment such that there is a one-to-one mapping between the two sets of regions. In this way is a certain supervenience guaranteed between mental contents and neural vehicles: there should be no mental (content) differences without physical (vehicular) differences. We do not want to attribute multiple contents if the organism is not capable of distinguishing them. This is analogous

to the case of the representation of the past in our experiment. The chemosensory input carries information about both the present and the past, but the feed-forward networks are incapable of distinguishing present from past. The attribution of contents to an organism is an attempt to portray the world as it is carved up by the creature's point of view: elements of the world that the creature cannot distinguish cannot make a difference discernable from the creature's point of view.

Based on the above sorts of arguments, we draw the following conclusions about the nature of representation, at least as it applies to the simplest cases of creatures behaving intelligently in virtue of possessing mental representations. Attributions of representations to organisms are not simply heuristics to be abandoned later when better ways of explaining behavior are discovered. They are attributions of real properties of organisms and real relations between organisms and their environments. The representations attributed are states of the nervous systems of the creatures that represent environmental (and bodily) states in virtue of carrying information about those and a requirement on the acquisition by the organism of such states is that the states enter into isomorphism relations between neural and other structures. The account so far is just a sketch and much will be revised and rejected soon.

Another feature of these representations worth noting for understanding points of view is their egocentricity. The distances represented in sensory and memory states are distances in an egocentric frame of reference.

There is much that is unsatisfactory with the account so far and key problems may be highlighted by considering the philosophically vexing problem of the representation of nonexistent objects. The representations of things that do not exist: gold mountains, square

circles, constitute one of the largest problem cases that inspire philosophers to worry about representation. It might even be framed as an objection to our view that our account of representation, couched in terms of information and isomorphism, cannot account for the representation of inexistents, since if something doesn't exist something else can neither carry information about it nor be isomorphic to it.

Perhaps the best thing to say at this point is that the roles information and isomorphism are playing in the above sorts of explanations are ones that highlight not representational content, but truth. The worms have such-and-such successful behaviors because they have sensory stand memory states that are *true*. The information and isomorphism relations track truth, not representational content. This leaves open to identify representational content with something else, then. Something other than something constituted by informational and isomorphism relations. One plausible candidate for content is constituted by the relations that define a representation's place in an internal structure.