CHAPTER 5

Evolving artificial minds and brains

Pete Mandik, Mike Collins and Alex Vereschagin

We explicate representational content by addressing how representations that explain intelligent behavior might be acquired through processes of Darwinian evolution. We present the results of computer simulations of evolved neural network controllers and discuss the similarity of the simulations to real-world examples of neural network control of animal behavior. We argue that focusing on the simplest cases of evolved intelligent behavior, in both simulated and real organisms, reveals that evolved representations must carry information about the creature's environments and further can do so only if their neural states are appropriately isomorphic to environmental states. Further, these informational and isomorphism relations are what are tracked by content attributions in folk-psychological and cognitive scientific explanations of these intelligent behaviors.

1 Introduction

Many kinds of explanations of intelligent behavior make reference to mental representations, that is, they explain an organism's ability to behave intelligently in virtue of an organism's having mental representations. The existence of such explanations, "representational explanations" for short, raises many questions, of which two will be the focus of this paper. The first is the question of whether representational explanations of intelligent behavior are the best explanations of intelligent behavior or if we might instead do better with explanations that make no reference to mental representations. We will argue that we can do no better than representational explanations. The second question that arises is the question of what representational explanations are referring to when they refer to representations. What *are* representations? We demand not just an account of what representations are, but additionally we demand an account that explains how representations can be the sorts of things that help explain intelligent behavior. We will sketch such an account. The goal of this paper then, is twofold: to argue that we need representations to explain intelligent

behavior and to sketch an account of what sorts of things representations must be if they are to explain intelligent behavior.

Several opponents of representational explanations have built their cases by starting with the simplest examples of intelligent behavior and attempting to demonstrate that in such examples, no representations are to be found and thus, no representations need be referred to in order to explain the behaviors at hand. This is the strategy followed, for example, by roboticists and artificial intelligence researchers such as Brooks (1991) and Beer (1990) in their arguments for the possibility of intelligence without representation. We will employ a similar strategy but we will be drawing different conclusions. We will examine some of the simplest cases of intelligent behavior and demonstrate that in these cases the behavior at hand is best explained in terms of representations. Further, our account of representations will be fully realist and reductive. To say that the account is realist is to say that the attributions aren't purely instrumental ways of speaking *as if* the creatures had representations. It is instead to pick out states of creatures that would be there independently of our speaking of them. To say of our account that it is reductive, we will be identifying representational states in ways that are straightforwardly explicable in terms of states of creatures' nervous systems and relations between their neural states and environmental states.

One way to examine the simplest examples of intelligent behavior is to examine the simplest examples of organisms that behave intelligently. This strategy confers the following advantage. The simpler the creature the easier it will be to keep track of the creature's internal structures, the structures of the creature's environment, and the relations between the two kinds of structure 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.

Our motive for caring about the evolutionary background of the simplest cognitive behaviors emerges from the following presumptions. We presume, and are 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. We 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 abiogeneis 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.

We 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 relations that constitute representation and the roles representations play in adaptive intelligent behaviors.

The structure of the rest of the paper is as follows. First we will briefly examine a few cases in which representations are invoked to explain the intelligent behaviors of humans and non-human animals. The goal here will not be to extract a definition of representation from these examples but instead to only note a few key features of the roles representations play in such explanations. Formulating a definition of representation is a goal to be achieved (or at least approximated) toward the end of the paper and not a presupposition to be made at its outset. Following the examination of these sample explanations, we will describe the basic intelligent behavior of positive chemotaxis and highlight the ways in which the problem that chemotaxis poses for organisms can be solved in a variety of ways involving representations. Next we describe mathematical and computer models of positive chemotaxis. The models are informed by neuroanatomical and neurophysiological data from real animals. Finally we discuss what account of representation seems best supported by the models.

2 Mental representations in explanations of intelligent behavior

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 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.

It is worth noting that the power of representational explanation is not simply some story we tell ourselves and each other sustained by our own (possibly mistaken) views of ourselves. One way to appreciate the power of such explanations is to appreciate them in the context of explaining the behaviors of non-human animals. The literature is filled with such examples. We briefly mention just a few. Consider the impressive feats of maze learning exhibited by rats. A Morris water maze is filled with water rendered opaque to obscure a platform that will offer a rat a chance to rest without having to tread water. When placed in the maze for a first time, a rat will explore the area and eventually find the platform. When the rat is

returned to the starting position, the rat does not repeat the exploratory strategy but instead swims straight to the remembered location of the platform. Apparently, the perceptual inputs gained during the exploration were utilized to compute the straight-line path to the platform. The rat's behavior is thus explicable in terms of psychological states such as perceptions and memories and computations that operate over them. Much more detail can be given, to be sure, but for now our main concern is only to call these sorts of explanation to the reader's attention. Much more detail concerning, for instance, the neural underpinnings of perception, memory, and computation, will be supplied later. Gallistel (1990: 2) describes another such example:

> Every day two naturalists go out to a pond where some ducks are overwintering and station themselves about 30 yards apart. Each carries a sack of bread chunks. Each day a randomly chosen one of the naturalists throws a chunk every 5 seconds; the other throws every 10 seconds. After a few days experience with this drill, the ducks divide themselves in proportion to the throwing rates; within 1 minute after the onset of throwing, there are twice as many ducks in front of the naturalist that throws at twice the rate of the other. One day, however, the slower thrower throws chunks twice as big. At first the ducks distribute themselves two to one in favor of the faster thrower, but within 5 minutes they are divided fifty-fifty between the two "foraging patches." … *Ducks and other foraging animals can represent rates of return, the number of items per unit time multiplied by the average size of an item*. (emphasis ours)

In both the cases of the rats and the ducks, the ultimate explanation called for is going to require mention of some relatively subtle mechanisms inside of the animals that are sensitive to properties of the environment. To get a feel for what might be called for, contrast the way in which we would explain, on the one hand, the movements of the rat toward the platform or the duck toward the bread and, on the other hand, a rock falling toward the earth. The rock's movement is explained by a direct appeal to a fundamental force of nature that constitutes the attraction between the respective masses of the earth and the rock. Such a direct appeal to a fundamental force will not explain the rat's movement to the platform. This is not to say, of course, that something non-physical is transpiring between the rat and the platform. There is of course energy flowing between the two that impacts the rat in ways that ultimately explain its behavior. But unlike the case of the rock, the transference of energy from platform to rat will only have an impact on the rat's behavior insofar as the rat is able to transduce the information carried by that energy into a code that can be utilized by information processing mechanisms in its central nervous system. Such mechanisms will be able to store information in the form of encoded memories and make comparisons between encoded memories and current sensory input to compute a course of action toward

a goal state. Going into further detail of how the nervous system of an animal might encode such information and perform such computations can get quite complicated. Before proceeding it will be useful to turn our attention toward nervous systems much simpler than those of vertebrates.

3 Modeling the simplest forms of intelligence

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 of 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 in 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, after walking for a while 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 (Ferrée & Lockery 1999). First off, the worms are able to effect chemotaxis even when their tail sensors are removed. Second, the 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 effectively 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 Ferrée and Lockery (1999: 263–277) and Dunn et al. (2004). Ferrée 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. Ferrée 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. Ferrée 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, Ferrée 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", Ferrée 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 adding the first order but omitting the zeroth order term.

Ferrée 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" (Ferrée & Lockery 1999: 276). 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 account of the mechanism, by "mechanism" we intend it in the sense of Craver (2001: 58): "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" (See also Craver & Darden 2001; Machamer, Darden & Craver 2000; Bechtel & 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 mechanisms underlying chemotaxis are. We are not in a comparable position yet with Ferrée 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: 381) 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.

Lockery and his colleagues are not blind to this sort of shortcoming. In a subsequent publication Dunn et al. (2004: 138) write "The chemosensory neurons re-

sponsible 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 were 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 (e.g. 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 hyopothesis, 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.

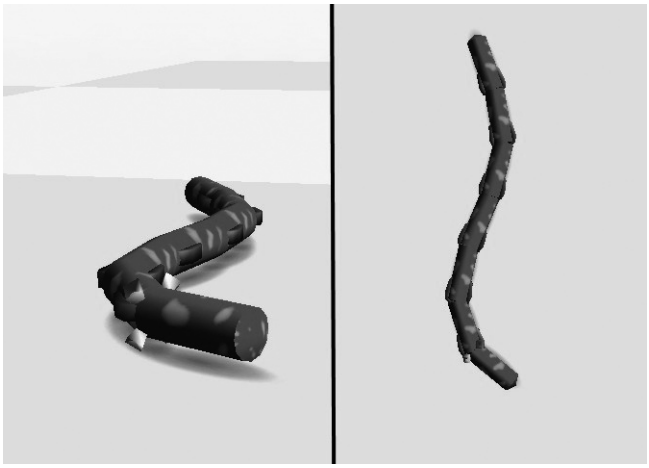

Figure 1. Synthetic C. Elegans. On the left, front view. On the right, top view

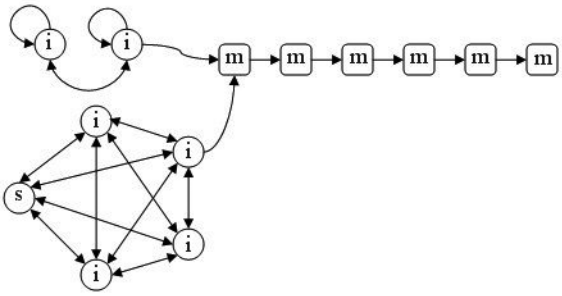

Figure 2. Neural network for the synthetic C. Elegans. Neurons include one sensor (s) and several motor neurons (m) and interneurons (i). Single-headed arrows indicate flow of information from one neuron to the next. A double-headed arrow between two neurons indicates both a feed-forward and a feedback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

In our simulations the initial morphologies and network topologies were set by hand. The connection weights, however, were 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 lifetime distance. 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 three 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 three 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 involved 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 were 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.

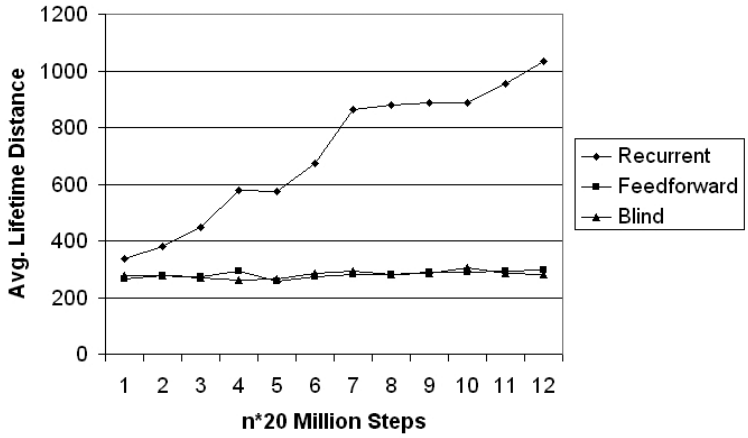

Figure 3. Results of the experiment comparing recurrent, feed-forward, and blind networks in an evolutionary simulation of chemotaxis

4 What the representations are in the models

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 and supported by the simulation experiments. Further, we are in a position with respect to these models to make some remarks about what the representations are and what relations obtain that determine 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, and the motor representations are states of activation in neurons that output to muscles. In each case, 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.

The question arises of what the relation is between representation and the 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 (Mandik 1999, 2005).

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 exists between a set of temperatures and a set of heights 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 to 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 isomorphsims between structures defined by ensembles of neural states and structures defined by ensembles of environmental states. In brief, in order to have the 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 only in terms of sets of elements and 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 is 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 or heavier players. Nonetheless, they still satisfy the requirements for isomorphism, since the mapping is structure preserving. Similarly, even if the information bearing states of a nervous system constitute a set of nominal representations of environmental states, they would nonetheless satisfy the requirements for isomorphism.

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 slightly higher level of neural activation can be used to carry information about a slightly 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: 356–537) 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.

(For an argument similar to Cummins' see also Churchland 2001: 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.)

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.

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-mental representations such as scientific representation (Suarez 2003) and the representations pertinent to measurement theory (Matthews 1994; Suppes & 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 relations 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. Similarly, we want to say 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. Thus there 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.

We close by briefly mentioning what the above account of representation in very simple systems might possibly say about the philosophically vexing problem of the representation of inexistent objects. The problem of representations of things that do not exist – gold mountains, square circles, etc. – constitutes one of the largest problems that inspire philosophers to worry about representation. It might even be framed as an objection to our view: Our account of representation, couched in terms of information and isomorphism, cannot account for the representation of inexistent objects, since if something doesn't exist something else can neither carry information about it nor be isomorphic to it. We have several brief remarks to make on this topic.

The first remark is that our primary concern is to give an account of how representations might underwrite certain kinds of explanation and it is unclear that representations of inexistent objects play roles in explanations in virtue of their contents and not simply in virtue of their vehicular properties. An underappreciated point is that we may very well be wrong when we think something has con-

tent. It may very well be that any appearances to the contrary from the first person point of view are explicable in terms of the first person indistinguishability of vehicles that have content and vehicles that do not. We suggest, then, that certain seeming representations, namely, so-called representations of things that do not exist, actually have no content. If the content of a representation is identical to the thing it represents, then a representation of a thing that does not exist is a representation with a content that does not exist. If nothing exists that is identical to the representation's content, then the representation has no content. Whatever role such representations play in explanations of behavior must, then, be due to their vehicular properties.

This line of thought derives a nihilism about content from a relatively dismissive view of inexistent objects. Some philosophers, however, may be much more tolerant of things such as non-actual possible worlds. Our response to such philosophers is that insofar as there may be a sense in which things that do not exist in the actual world nonetheless exist, then there may be a sense in which things in non-actual worlds are able to enter into the requisite information and isomorphism relations with neural representations in the actual world. (See, for example, Swoyer 1991 for a discussion of isomorphism based representations and non-actual possible worlds.)

We close, then, with a final remark on the topic of the representation of inexistent objects. Whatever the status of the representation of inexistent objects, it is safe to say that such representations are irrelevant in the explanations of the simplest cases of cognition. Certainly, the representation of a thing that does not exist cannot be the most basic case of representation. Restricting our attention to the simplest cases of representation, we see that the simplest cases of things that can be represented are things that not only actually exist, but also actually enter into relations of information and isomorphism with neural states.

Acknowledgements

This work was supported in part by grants to Pete Mandik from the National Endowment for the Humanities and James S. McDonnell Foundation Project in Philosophy and the Neurosciences. Pete Mandik is grateful for feedback from members of audiences of presentations of this material at the 2004 International Language and Cognition Conference in Coffs Harbour, Australia; the City University of New York Graduate Center Cognitive Science Symposium and Discussion Group; and the Neurophilosophy: The State of the Art conference at Caltech.

Mike Collins is grateful to audiences at the City University of New York Graduate Center Cognitive Science Symposium and Discussion Group and the Fall 2004 meeting of the New Jersey Regional Philosophical Association at Felician College.

References

Akins, Kathleen. 2001. Of sensory systems and the 'aboutness' of mental states. In William Bechtel, Pete Mandik, Jennifer Mundale & Robert S. Stufflebeam (eds.), *Philosophy and the neurosciences: A reader*, 369–394. Oxford: Blackwell.

Bechtel, Willam & Robert Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Beer, Randall. 1990. *Intelligence as adaptive behavior*. San Diego, CA: Academic Press.

Brooks, Rodney. 1991. Intelligence without representation. *Artificial Intelligence* 47. 139–159.

Churchland, Paul. 2001. Neurosemantics: On the mapping of minds and the portrayal of worlds. In K. E. White (ed.), *The emergence of mind*. Milano: Fondazione Carlo Elba.

Churchland, Paul. 2002. Catching consciousness in a recurrent net. In Andrew Brook & Don Ross (eds.), *Daniel Dennett: Contemporary philosophy in focus*, 64–80. Cambridge: CUP.

Clark, Andy & Josefa Toribio. 1994. Doing without representing? *Synthese* 101. 401–431.

Craver, Carl. 2001. Role functions, mechanisms and hierarchy. *Philosophy of Science* 68. 31–55.

Craver, Carl & Lindley Darden. 2001. Discovering mechanisms in neurobiology: The case of spatial memory. In Peter K. Machamer, Rick Grush & Peter McLaughlin (eds.), *Theory and method in neuroscience*, 112–137. Pittsburgh, PA: University of Pittsburgh Press.

Cummins, Robert. 1996. *Representations, targets, and attitudes*. Cambridge, MA: The MIT Press.

Cummins, Robert. 1997. The LOT of the causal theory of mental content. *Journal of Philosophy* 94. 535–542.

Dunn, Nathan A., Shawn R. Lockery, Jonathan T. Pierce-Shimomura & John S. Conery. 2004. A neural network model of chemotaxis predicts functions of synaptic connections in the nematode *Caenorhabditis Elegans*. *Journal of Computational Neuroscience* 17(2). 137–147.

Ferrée, Thomas C. & Shawn R. Lockery. 1999. Computational rules for chemotaxis in the nematode *C. Elegans*. *Journal of Computational Neuroscience* 6. 263–277.

Gallistel, Charles R. 1990. *The organization of learning*. Cambridge, MA: The MIT Press.

Komosinski, Maciej. 2000. The world of Framsticks: Simulation, evolution, interaction. *International Conference on Virtual Worlds* 2. 214–224.

Kulvicki, John. 2004. Isomorphism in information carrying systems. *Pacific Philosophical Quarterly* 85. 380–395.

Lloyd, Dan. 2003. *Radiant cool: A novel theory of consciousness*. Cambridge, MA: The MIT Press.

Machamer, Peter, Lindley Darden & Carl Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67. 1–25.

Mandik, Pete. 1999. Qualia, space, and control. *Philosophical Psychology* 12(1). 47–60.

Mandik, Pete. 2002. Synthetic neuroethology. *Metaphilosophy* 33(1–2). 11–29.

Mandik, Pete. 2003. Varieties of representation in evolved and embodied neural networks. *Biology and Philosophy* 18(1). 95–130.

Mandik, Pete. 2005. Action oriented representation. In Andrew Brook & Kathleen Akins (eds.), *Cognition and the brain: The philosophy and neuroscience movement*. Cambridge: CUP.

Matthews, Robert. 1994. The measure of mind. *Mind* 103. 131–146.

Suarez, Mauricio. 2003. Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science* 17(3). 225–244.

Suppes, Patrick & Joseph L. Zinnes. 1965. Basic measurement theory. In R. Duncan Luce, Robert R. Bush & Eugene Galanter (eds.), *Handbook of mathematical psychology*, 1–76. New York, NY: John Wiley and Sons.

Swoyer, Chris. 1991. Structural representation and surrogative reasoning. *Synthese* 87. 449–508.