For Whom the Bell Tolls? The Roles of Representation and Computation in the Study of Situated Agents*

Michael Wheeler
School of Cognitive and Computing Sciences,
University of Sussex, Brighton BN1 9QH, U.K.
Telephone:+44 273 678524
Fax:+44 273 671320
E-Mail:michaelw@cogs.susx.ac.uk

Abstract

Orthodox cognitive science claims that situated (world-embedded) activity can be explained as the outcome of in-the-head manipulations of representations by computational information processing mechanisms. But, in the field of Artificial Life, research into adaptive behaviour questions the primacy of the mainstream explanatory framework. This paper argues that such doubts are well-founded. Classical A.I. encountered fundamental problems in moving from toy worlds to dynamic unconstrained environments. I draw on work in behaviour-based robotics to suggest that such difficulties are plausibly viewed as artefacts of the representational/computational architecture assumed in the classical paradigm. And merely moving into connectionism cannot save the received orthodoxy. If we adopt the perspective according to which neural networks are most naturally conceptualized as dynamical systems, it becomes appropriate to treat such networks as computational devices only if the network-dynamics are deliberately restricted. A different explanatory framework is required once artificial neural networks are developed both to exhibit dynamical profiles comparable to those displayed by biological neural networks, and to play the same adaptive role as biological networks, i.e., to function as the control systems for complete situated agents. I close by describing an example of a dynamical systems explanation of situated activity.

1 Two Dogmas of Cognitive Science

The orthodox view in cognitive science has its roots in the representational theory of mind — the empirical hypothesis which claims that the crucial aspects of cognition involve the processing of semantically interpretable internal states which function to encode objective states of an external world. This internal-representation-hypothesis is easily linked to the complementary thought that cognition is fundamentally computational, because the existence of 'in-the-head' computational information processing mechanisms would seem to provide some hope of an explanation as to how a physical system could realize a representational system. Stated thus, these are extremely broad commitments, since what is meant exactly by 'representation' or 'computation' can differ from theory to theory, and

 Parts of this paper are to appear (as *From Activation to Activity: Representation, Computation and the Dynamics of Neural Network Control Systems*) in *Artificial Intelligence and Simulation of Behaviour Quarterly — special issue on Artificial Life.*

from model to model. During the course of this paper, I shall describe some alternatives. But notice that the received explanatory framework, as I have characterized it, covers both classical theories and (most) connectionism.

Once an interesting notion of 'causally efficacious internal state' plays some role in the explanatory story, representational/computational accounts seem to find a foothold. But whilst it may seem just *obvious* to many researchers that this orthodox framework provides the appropriate explanatory tools for the scientific explanation of the relevant behaviour, there is a line of research in Artificial Life which indicates that the priority usually accorded to the concepts of 'representation' and 'computation' is far from guaranteed. Indeed the time to ring the bell signalling the end of the existing orthodoxy may well be nigh. Exploring just such a possibility is the purpose of this paper.

2 Artificial Life and Situated Agents

The amorphous nature of the set of interests and approaches brought together under the umbrella-term 'Artificial Life' (A-Life) — from models of RNA replication and sensory-motor activity to collective intelligence and population dynamics — makes defining the scope of the field tricky, to say the least. I shall concentrate on those areas of research which have a direct bearing on the argument of this paper.

In A-Life an *autonomous agent* is a fully integrated, self-controlling, adaptive system which, while in continuous long-term interaction with its environment, actively behaves so as to achieve certain goals. So for a system to be an autonomous agent, it must exhibit *adaptive behaviour*, behaviour which increases the chances that that system can survive in a noisy, dynamic, uncertain environment. We should identify a system as an adaptive system only in those cases where it is useful to attribute survival-based purpose and purposes to that system. So rivers don't count as adaptive systems, but moths do. Naturally-occurring adaptive behaviour is the result of evolutionarily determined pressures on the survival and reproduction prospects of embodied creatures. Hence the class of naturally-occurring autonomous agents includes humans, non-human mammals, fish and insects.[1]

On evolutionary grounds, it seems reasonable to hypothesize that human linguistic competence and deliberative thought are overlays on a prior (and, in terms of survival, more fundamental) capacity for adaptive behaviour. Under the influence of this sort of thought, the A-Life-orientated search for an understanding of intelligence starts not with the sort of reasoning capacities possessed by humans, but with the adaptive behaviour of simpler, although whole, situated agents that perceive and act. The A-Life methodology is to develop complete control systems for *artificial autonomous agents* — often called *animats* [28]. Animats can be real autonomous robots with actual sensory-motor mechanisms, or simulated agents in interaction with simulated environments. The aim of such work is not simply to produce useful robots which exhibit robust behaviour in uncertain environments. The goal is to increase our understanding of the mechanisms underlying adaptive behaviour, through the synthesis and analysis of artefacts.[2]

[1] As with most (all?) definitions of concepts, there are potential problem cases. By the definitions offered here, some plants might count not only as adaptive systems, but also as autonomous agents. I shall just stipulate that, in the context of this paper, the class of autonomous agents excludes plants. To me such a move is intuitively correct; but I accept that some may find it more than a a little arbitrary.

[2] The fact that the behaviour of a system is simulated on a computer does not automatically mean that that system is best explained in computational terms. We can use a computer to simulate the behaviour of a fluid, without concluding that the fluid is computing what to do. Similarly we may use representational

A-Lifers tend to adopt a standpoint according to which cognition should be seen as an adaptive phenomenon. On this view (which I endorse) we can make sense of a cognitive system as the control system for an autonomous agent because we can make prior sense of that creature's environmentally embedded behaviour — its *situated activity* — as adaptive behaviour. Two immediate implications of this general way of conceptualizing cognition are that cognitive science itself should begin as the science of situated activity and that the fundamental properties of naturally-occurring cognition (and not just its 'mechanical realization') can be investigated by the biological sciences (including neurophysiology, ethology, behavioural ecology and evolutionary theory). This second implication is in harmony with the fact that in A-Life, biological constraints are not (as they are in most orthodox cognitive science — see later sections) thought of as 'mere' implementational details. On the contrary, biological factors are considered crucial to an understanding of the adaptive phenomenon in question.

3 A Shakey Start

In the classical approach to Artificial Intelligence (A.I.), a 'representation' is thought of as either an atomic symbol or a complex molecular structure constructed through the systematic recombination of simpler symbolic elements according to syntactic rules. The meaning of a molecular representation is a function of the meanings of the constituent symbols plus the syntactic structure of the complex formula. In other words, classical representational systems feature a combinatorial syntax and semantics. With such a structure to the representations, the computational principles by which those representations are manipulated or transformed can be defined over the structural properties of those representations, the 'computations' being the ordered steps through which the manipulations and transformations are achieved.[3]

Classical A.I. embraces the principles of *homuncular decomposition,* i.e., the view that we can compartmentalize a system into a hierarchy of specialized sub-systems that (i) solve particular sub-tasks by manipulating and/or transforming representations through computations and (ii) communicate the computed outputs to each other by passing representations. It is important to stress that homuncular talk of 'little people' in the head is strictly metaphorical. The commitments are to the actual existence of internal structures that the external observer can usefully interpret as information-bearing representational states, and to the actual existence of internal modules that the external observer can usefully interpret as carrying out the manipulation, sending and receiving of representational tokens in order to realize some overall input-output mapping.

In sharp contrast to the fundamental tenets of most work in A-Life, classical A.I. has concentrated on abstracted sub-domains of human cognition (such as natural language processing or formal reasoning) with no master-plan for how to integrate all the different specialist modules either with one another, or with sensing and action, to create a complete, intelligent agent. (Even work in computer-vision has tended to concentrate on scene analysis and to cut itself off from questions of ongoing activity in a world.) Furthermore, the classical assumption has been that the inevitably messy and complicated business of achieving real-time interaction with an environment is essentially an implementation headache to be overcome by the hardware department. Under these circumstances, it is relatively unsurprising that robotics rarely crept into the spotlight. But my arguments in

frameworks (such as mathematics) in our modelling processes without concluding that the systems under investigation necessarily use those representations in order to achieve the observed behaviour [4].

[3] Newell and Simon [23] present a full statement of this view..

this paper are concerned directly with the control systems required by situated autonomous agents. So it is instructive to take note of what happened when classical A.I. actually concerned itself with robots.

Classical robots (e.g., *Shakey* [24]) featured control systems designed according to the following principles (dubbed "decomposition by function" by Brooks [7, 8]). A perception-module constructs a symbolic (conceptual-level) description of the external world. This world-model is then delivered to a central system made up of sub-modules for specialized sub-problems such as reasoning and planning. These sub-modules manipulate the representations in accordance with certain computational algorithms, and then output a further symbolic description (this time of the desired actions) to which the action-mechanisms then respond. Such organizational principles clearly respect the concept of homuncularity.

Lurking behind the classical methodology is a crucial premiss to the effect that, even given accuracy problems resulting from noisy or drifting sensory-motor mechanisms, it is still *possible* to build an adequate, stored world-model that can be manipulated in real-time. This is required so that, for the purposes of planning action, operating in an actual world can be ignored in favour of the internal representations. But as A-Lifers (and others) have observed, once an autonomous agent's domain of activity is a dynamically changing, uncertain environment, a commitment to maintaining an accurate internal world-model could well be a devastating error. The problem amounts to an explosion in the demands placed on representational and computational resources. This makes the problem intractable on the time-scales relevant to the realization of adaptive behaviour, a fact which would signal the untimely end of many a predator-threatened animal.

It is here that one of the most formidable hurdles to confront orthodox artificial intelligence comes to the fore — the notorious *frame problem.* In its strictest form, this is the problem of characterizing the aspects of a state that are not changed by an action. However, it has come to be used to name a family of related problems to do with update and relevancy. The basic question can be posed like this: how, given particular sets of circumstances, goals and actions, does an autonomous agent come to respond to those state-changes in its world which really matter, whilst ignoring those which are irrelevant? Having relevancy heuristics just won't do; how do the processing mechanisms access just those relevancy heuristics which are relevant? An infinite regress threatens.

Our model-building classical robot must meet the challenge presented by the frame problem, because, to act effectively, that robot has to keep its internal world model in step with its external environment. In toy worlds, the designer can 'overpower' those aspects of the frame problem that arise, either by taking comprehensive account of the effects of every action or change, or by working on the assumption that nothing changes in a scenario unless it is explicitly said to change by some operator-definition. This explains why the frame problem was nothing more than a nuisance in the 'blocks-world' simulations popular in certain stages of the A.I.-enterprise. (Blocks-worlds were artificially restricted task-domains in which programs confronted toy problems, and in which the human designer prescribed the semantics of the environmental properties and relations of importance.)

Given this, it is a telling observation that robots instantiating the classical principles of organization tended to be highly dependent for their performance on the fact that their operational environments were carefully engineered to suit the robots' processing strategies. For example, in the case of *Shakey*, the robot's environment consisted of rooms sparsely populated by static blocks. These blocks were painted different colours on different planar surfaces to facilitate edge detection in a visual image. Moreover, *Shakey's* environment was essentially static. A 'demon' was occasionally allowed to alter the position of some of the obstacles when *Shakey* wasn't looking, but this hardly makes

the environment dynamic in any ordinary sense. And notice that the human designer plays the same role in the case of the environmental properties and relations to be recovered by the classical robots as she does in the case of blocks-worlds. That is why the designers of *Shakey* could adopt the second of the identified toy-world 'solutions' to the frame problem. (One response to this sort of observation would cite the possible role of learning algorithms in improving the adequacy of the robot's representations. But as long as the semantics of the task-domain are carefully prescribed by the human-designer, and the robot's job is to build an objective internal model of the properties and relations of its environment by using the designer's pre-specified semantic primitives, we are still in the blocks-world — whether or not learning is part of the process.)

So the evidence suggests that it is possible to adopt the sort of strategies deployed in classical robots only in those cases where the environment is specially, and artificially, controlled. And things get worse (for the classicist). When an engineer approaches the task of designing a system to solve a complex problem, the standard tactic is to decompose the problem so that it can be collectively surmounted by simpler, communicating subsystems with well-defined functions and interfaces. In general, then, engineers work with well-specified problems, and engineering solutions reflect the designer's functional conceptualization of the problem. Such a methodology is deeply entrenched in computational engineering, in which, as we have seen, functionally specified modules — homunculi — carry out well-defined computations and communicate with each other via representations.

But there is reason to think that the problem of synthesizing environmentally embedded adaptive behaviour is not well-defined enough for the traditional human-intervention in the input-output loop generally to be profitable. For animals, the primary adaptive goal is to survive long enough to reproduce. In a noisy, dynamic, and possibly hostile environment, the constraints on achieving this goal are not only inherently difficult to specify but, because of the existence of coevolutionary situations, where adaptations by one species effectively alter the environment of another species, the problem itself is subject to evolutionary change. If artificial autonomous agents are embedded in similarly dynamic and uncertain environments, then the relevant constraints will also be difficult to specify and unavoidably open-ended. Moreover, natural evolution merely retains the designs of those creatures which consistently survive long enough to reproduce. The only constraint on the agent's internal dynamics is that they allow the system to achieve the required adaptive behaviour. In nature there is no assumption to the effect that the organization of the agent's control system must embody the sort of computational-style decomposition traditionally favoured by human designers.

4 Breaking the Mould

For various reasons, the field of *behaviour-based robotics* (e.g., [7, 8, 14]) has become allied with the A-Life movement. The behaviour-based approach advocates highly reactive control architectures, with no central reasoning systems, no manipulable symbolic representations, and radically decentralized processing. The idea is that individual behaviour-producing systems, called 'layers', are individually capable of — and generally responsible for — connecting the robot's sensing and action in the context of, and in order to achieve, some ecologically relevant behaviour. Then, starting with layers which achieve simpler behaviours such as 'avoid objects' and 'explore,' layers are added, one at a time, to a debugged, working robot, so that overall behavioural competence increases incrementally. The layers run in parallel, affecting each other only by means of suppression or inhibition mechanisms.

Behavioural decomposition is clearly at odds with the classical picture. The principles of homuncularity do not apply and there is no central locus of reasoning and control. In fact, the process of attempting to build a centrally stored, 'objective' world model is rejected as constituting a positive hindrance to real-time activity in a messy environment. In its place is a view according to which a situated agent should operate by continuously referring to its sensors as opposed to some internal representation. A world is a source of surprises, but it is also a source of informational continuity through its ongoing history.[4] However the aim is not to reject outright any form of representation. In fact, there may well be some representational interpretation of *certain* individual layers. For example, Franceschini *et al.* [15] implement a two-layered behaviour-based architecture in which a 'goal pursuit' layer runs in parallel with an obstacle avoidance layer. The goal pursuit layer functions by constantly defining a robot-egocentric deictic map of obstacles in polar coordinates, in relation to the instantaneous direction in which the robot is heading. The map is not an objective representation which is stored, recalled and updated; rather it is agent-centred and dynamically created as the robot moves through its environment. In this approach, the classical separation of data-structure and computation is not present; and a 'representation' is a decentralized, non-manipulable, essentially *active* structure, used in the context of a specific behaviour. All of this is in contrast to the all-purpose, task-independent, object-centred world-models favoured in the classical paradigm.

How does this relate to the sorts of difficulties faced by the classical approach to situated action? Each layer in a behaviour-based control system is closely coupled to the robot's environment along what might be called a 'channel of ecological significance' which connects sensing to action in the context of the specific adaptive behaviour. Thus the paradigm seemingly by-passes the computationally intractable problem of maintaining centrally-stored objective representations. In addition, certain cognitive theorists (such as Churchland [9]) have argued that the frame problem itself is an artefact of the classical form of representation. So it is conceivable that the fundamental commitments of the behaviour-based approach (including the parsimonious deployment of deictic representations) will allow robots to side-step (temporarily at least) the debilitating effects of the frame problem. Indeed, the approach has produced robots that have performed tasks — generally with reasonable competence — in uncertain, dynamic environments. For example, *Herbert* [14], a robot featuring a fourteen-layer behaviour-based architecture, moved autonomously around the cluttered, people-populated halls of M.I.T., stealing empty soda-pop-cans and depositing them in a central bin.

The message of this section is that, in the battle to achieve adaptive behaviour in artefacts, progress has been made by A-Life roboticists who recognize that many adaptive problems are solved primarily through the dynamics of the interaction between the agent and its environment, and not by the construction and manipulation of objective representations. If this is the right way to go, then the very notion of a 'representation' must undergo what, on even the most conservative estimate, amounts to a fundamental transformation. But a more radical conclusion may be beckoning.

[4] As Boden [6] reminds us, the observation that the best source of information about the world is the world itself was made by some researchers in orthodox cognitive science. Unfortunately it was no sooner made than forgotten by most of the field.

5 Thinking about Networks

In the natural world, sophisticated situated activity arises through the incremental evolutionary development of biological neural networks as the control systems for animals. So it seems likely that an artificial neural network *of some description* will provide a profitable basic structure for an animat control system. But what sort of networks should be the focus of attention? And what implications does the answer to that question have for the explanatory roles of representations and computations? The thought here is that whilst the received explanatory framework may well be appropriate for the sorts of models and scenarios usually studied by connectionists, such apparent domestic harmony may mask a rather unstable marriage of convenience. This instability becomes manifest once we face the sorts of issues which dominate A-Life.

The overwhelming majority of connectionists conceptualize neural networks as computational devices which, by calculating outputs from inputs, effect transformations between representations. The most commonly championed sense of 'connectionist representation' refers to a pattern of activation distributed over a network of units, where an individual unit may have no isolable representational interpretation. The global computations — the transformations between input and output representations — occur as spreading activation patterns on the basis of local computations by individual units. These local computations are determined, in part, by the values of the connection strengths.[5]

Any attempt to articulate a radically different view of neural networks depends on there being a conceptual framework for understanding network-activity which does not assume representations and computations as theoretical primitives. Fortunately such a framework is readily available. A number of connectionists concerned with cognitive modelling have already suggested that *dynamical systems theory* is a natural explanatory language with which to describe the behaviour of neural networks.[6]

In essence, a dynamical system is any system for which, potentially at least, we have a rigorous analysis of the way it evolves over time. Formally, this is any system for which we can provide (1) a finite number of state variables which, given the interests of the observer, adequately capture the state of the system at a given time, and (2) a set of state space evolution equations describing how the values of those variables change with time. Given (1), we can produce a geometric model of the set of all possible states of the system called the *state space* of that system. A state space has as many dimensions as there are state variables for the system, and each possible state of the system is represented by a single point in that space. Given (2), and some initial conditions — a point in the state space — subsequent changes in the state of the system can be plotted as a curve in the state space. Such a curve is called a *trajectory* of the system, and the set of all such trajectories is the system's *phase portrait*.

An *attractor* is a state of the system to which trajectories passing nearby tend to converge. Now consider some attractor P. The set of states such that, if the system is in one of those states, the system will evolve to the attractor P, is called P's *basin of attraction.* The trajectories which pass through points in the basin of attraction on the way to the attractor, but which do not lie on the attractor itself, are *transients* of the system. There may well be several attractors in a single state space, and these attractors

[5]The literature on connectionist representation is *huge.* Clark [10] provides clear coverage of the key issues.

[6]See, for example, [20, 26, 27]. Although these papers all describe neural networks as high-dimensional dynamical systems, there is no overall agreement as to the implications of such a view for the status of computational/representational forms of explanation.

may be of different types. For instance, *point* attractors are single points in the state space which represent constant solutions to the system, whilst *periodic* attractors represent oscillatory solutions. *Chaotic* attractors represent highly complex behaviours in which a dynamical system exhibits what is known as *sensitive dependence on initial conditions*. This means that if two different initial states are chosen, which are even a tiny distance apart, the two subsequent trajectories will diverge from each other very quickly. On average, the divergence will be exponential. These exponentially diverging trajectories remain bounded on the attractor without intersecting. So they fold back on themselves, creating an infinitely layered chaotic attractor. Such attractors are common in high-dimensional, non-linear systems.[7]

Computational systems (defined by reference to Turing machines) are dynamical systems. But the set of computational systems (so defined) fills one tiny corner of the space of possible dynamical systems [16, 27]. This tells us only that that the language of dynamical systems theory provides a more general conceptual framework than that on offer from the orthodox computational camp. In the present context, what we need to know is whether the dynamics of connectionist networks lie outside the space occupied by computational systems.

So, given the characterization of a connectionist representation as a distributed pattern of unit-activations, let us think about the activation-space dynamics of a standard connectionist network during its processing stage. Typically, a human introduces input data to the system. This places the network at some initial point in activation space — a state space with as many dimensions as there are units in the network, and where a point in that space is defined by the simultaneous activation values of each of those units. If the network has been trained successfully, this initial state will be in the basin of attraction of a point attractor which (under some suitable semantic interpretation) encodes the correct solution. The successive states of the network will trace out a transient of the system through activation space on the way to the point attractor where, upon arrival, the system will come to rest.

It seems quite natural to describe such network-dynamics in the language of the orthodox framework. Indeed someone impressed by the explanatory power of representations and computations should not feel unduly threatened by a picture according to which the processing of a network is conceptualized as a trajectory through activation space from an initial state to a fixed point attractor. The start and end points of the trajectory can be decoded as vectors of activation values which, in a more or less standard fashion, can be treated as input and output representations with semantic interpretations. (Sometimes the interpretation of interest has to be decoded from hidden unit activity patterns using statistical techniques such as cluster analysis. This does not affect the fundamental dynamical profile.) Notice also that there is no violation of the principles of homuncularity. Either the network itself is carrying out some functionally well-defined sub-task and so can be viewed as one homunculus among many, or (as Harvey [17] observes) individual layers within a multi-layered network are thought of as modules which communicate with each other by passing representations. The processing story on offer here seems essentially equivalent to — or interpretable as — a matter of computing outputs from inputs through the manipulation and communication of representations. This is all well and good; but why should the activation-space dynamics of artificial neural networks be restricted to unperturbed trajectories to point attractors? It is time for an important reminder.

It has become depressingly commonplace to find far too much being made of the biolog-

[7] For a friendly but thorough introduction to dynamical systems theory, see [1].

ical 'feel' of standard connectionist networks. Although the processing architecture of such orthodox networks may resemble the basic abstract structure of the brain, there are significant areas of divergence between standard connectionist networks and their biological relations. For instance, certain restrictions often placed on connectivity in connectionist networks (e.g. feed-forward activation passes or symmetrical connections) are, generally speaking, not reflected in the real nervous systems; biological networks are inherently noisy; there is rarely any real correspondence between connectionist units and biological neurons (the latter are far more complex); and, unlike mainstream connectionist networks, biological networks typically feature many distinct types of neuron, a point which holds whether we are talking about the visual system of the fly [15] or the human brain [2]. The effect of these dissimilarities is that the intrinsic dynamics of standard connectionist networks are positively impoverished when compared to those exhibited by biological networks. To appreciate the complexity of neural dynamics, consider that one of the implications of Skarda and Freeman's [25] model of the neural dynamics underlying odor recognition and discrimination in rabbits is that it is the chaotic dynamics of the olfactory bulb which enable the system to add new odors to its repertoire of learned odors. A chaotic background state prevents convergence to previously learned neural response-states and allows the generation of new states. So the indication is that, without some reference to the relevant chaotic dynamics in the neural system, any neural account of such sensory recognition capacities would be incomplete (at best). It is surely likely that conclusions similar to those arrived at in the Skarda and Freeman study will apply to other sensory modalities and other sensory skills. The accompanying thought is that it must at least be plausible that network-dynamics more complex than fixed points will play a significant role in any explanatory paradigm which takes seriously the biological basis of cognition.

It is time to introduce a class of artificial neural networks which can do justice to the intuition that the sort of complex dynamics realized by biological neural networks may well be relevant both to synthesizing adaptive behaviour and to explaining cognition. Presumably the following architectural principles are not the *only* ones which would result in the style of complex dynamical behaviour I have in mind. However, these particular principles are suggestive because, to some extent, they capture certain styles of constraint (identified above) which are respected by real neural structures. Consider artificial neural networks that feature the following sorts of properties: deliberately introduced noise, continuous-time processing, real-valued time delays on the connections between the units, units with non-uniform activation functions, and connectivity which is not only both directionally unrestricted and highly recurrent, but also *not* necessarily subject to symmetry constraints. Artificial neural networks featuring some or all of these properties are championed in [4, 5, 13, 18, 21, 29]. These *dynamical neural networks* are capable of producing far richer intrinsic dynamics than those produced by mainstream connectionist systems.

To sum up this section: Orthodox connectionists draw on the gross abstract structure of extant biological networks, without paying due attention to the potentially relevant details of that structure. The standard restrictive architectural assumptions reflect the granting of minimal theoretical weight to the biological basis of naturally occurring cognition, a move which we can recognize as being made by classicist and connectionist alike. Whilst my intention is not to campaign on a 'no abstractions allowed' ticket (which would fly in the face of scientific method), certain key aspects of the dynamical profiles of biological neural networks may well be essential to our explanations of the corresponding cognitive capacities; i.e., the dynamics of artificial neural networks should reflect those of biological neural networks. What the 'key aspects' of the dynamics are is a matter to be decided by empirical research, not philosophical prejudice. "Fair enough" (perhaps I hear you say)

"but what does all this tell us about representational/computational ways of thinking?" We need take just one more A-Life-oriented step.

6 Situated Control Systems

Mainstream connectionists have tended to follow their classical cousins into abstracted sub-domains of cognition. But what happens when we confront not only neurobiological dynamics, but the dynamics of situatedness as well? A nervous system is a biological network that constitutes the basis of the control system for a situated agent. (Remember cybernetics.) So, in the study of artificial situated agents, it seems that we should investigate the properties and potential of dynamical neural networks as the control systems for animats.

Consider the A-Life paradigm of *Evolutionary Robotics,* where genetic algorithms (processes inspired by Darwinian natural selection) are used to develop dynamical neural network control systems for autonomous robots. (The work described here is by Cliff, Husbands and Harvey. See [12, 19] for fuller details.) Loosely speaking, the evolutionary methodology is to set up a way of encoding neural network architectures as genotypes, and then, starting with a randomly generated population of network-controllers, and some evaluation task, to implement a selection cycle such that more successful networks have a proportionally higher opportunity to contribute genetic material to subsequent generations. Over successive generations, better performing controllers are discovered. One general commitment of this work is that as few restrictions as possible are placed on the potential structure of the network. The evolutionary roboticist decides on the robot's immediate task, but endeavours to stay out of the business of how the robot's 'nervous system' should work in order to achieve the appropriate behaviour. For example, in the work just mentioned, the number of internal units, the number, directionality, and recurrency of the connections, and certain parameters of the visual system are placed under evolutionary control. (So the sensory-motor systems are considered to be part of the control system.) The job of artificial evolution is to tune the control-system-dynamics to the environment in such a way that the robot can complete the evaluation task. This approach has resulted in the artificial evolution of sensory-motor controllers which succeed in guiding robots in the performance of simple homing and target-tracking tasks, and which are adaptive in that they exhibit considerable robustness when tested on generalized versions of the specific tasks for which they were actually evolved.[8]

So the artificial control-networks are developed in such a way that there is no forced conformity to human-friendly styles of functional/homuncular decomposition, or, indeed, to the principles of behavioural decomposition. This is not to say that artificially evolved control systems will necessarily exhibit no form of modular decomposition. But no particular form of decomposition is assumed at the outset of the evolutionary process.[9] The 'hands-off' principle has a further effect. In standard connectionism, the human designer not only makes the key architectural decisions, but also intervenes directly in the semantics of the input-output loop. She uses a prior theory of the problem space (often informed by theories developed within classical approaches) to stipulate the semantic categories used by the network's input and output representations. These categories define the content

[8]Related approaches to the development of neural networks include, among others, [4, 5, 29]. In these studies a genetic algorithm is used to search a predefined finite space of possible network architectures. This is in contrast to the 'open-ended' approach adopted by Harvey, Cliff and Husbands.

[9]Significantly, there is evidence that some artificially evolved networks may submit to behavioural decomposition [21].

of the relevant input-output mappings. But in the evolutionary approach, the sensor and motor interfaces have no semantic interpretation, and all 'meaningful' interpretations of the robots' internal dynamics have to be settled 'after the event' so to speak, when the control network has been evolved. (Hence we witness the birth of computational neuroethology [3, 11].) So how should we go about explaining the environmentally embedded behaviour of situated agents who feature dynamical neural networks as control systems?

The dynamical systems approach to situated activity holds that an agent and its environment should be conceptualized as *coupled* dynamical systems.[10] The ongoing behaviour of a dynamical system is specified by the current state of the system and the evolution equations which govern how the system changes through time. (See section 5.) Certain values in a state space evolution equation specify quantities that affect the behaviour of the system without being affected in turn; these are called the parameters of the system. Any particular phase portrait will be defined relative to a specific set of parameter values. The crucial relation of *coupling* obtains when two separable dynamical systems are bound together in a mathematically describable way, such that, at any particular moment, the state of either system fixes the dynamics of the other system, in that some of the parameters of each system either become, or become functions of, some of the state variables of the other. Now consider two coupled dynamical systems, X and Y: X is said to *perturb* Y when changes in the state variables of X result in changes in the parameter-values determining the phase portrait of Y, thereby resulting in changes in that phase portrait. At some critical parameter-values a system may become *structurally unstable,* in that tiny changes in certain parameter-values may result in the immediate emergence of a qualitatively different phase portrait. These qualitative changes are called *bifurcation points.*

The sense in which one system affects the dynamics of another system through coupling is not to be equated with a relation according to which one system directly specifies the state of a second. By perturbing Y's dynamics, X is biasing the intrinsic possibilities for change already present in Y. So if we begin by thinking of an animal nervous system (its biological control system) as a non-coupled dynamical system, then we can conceptualize its intrinsic dynamics as generating a space of possible perturbations which the system can undergo as a result of coupling to an environment (cf. [22]). The relation between nervous system and environment is one of influence of dynamics rather than specification of state. Through sensory-motor activity, the dynamics of an animal's nervous system are continually perturbed in accordance with the adaptive agent-environment couplings 'discovered' by evolution. And that, according to the view advocated here, is the place to start in any study of situated cognition. So how does this dynamical perspective mesh with the received orthodoxy in cognitive science?

7 The Dynamics of Situated Activity

Recall that on the orthodox view, the fundamental basis of cognition is the representation-based recovery and use of objective environmental information. This 'representation-based recovery' is seemingly the specification, by the environment, of an essentially static structural state of the agent's central nervous system. One intuitively plausible way to cash out the realization of such states would be as point attractors in neural activation space, i.e., by mapping standard connectionist representations onto states of the nervous system, in such a way that the activity of one connectionist unit is hypothesized to correspond to the

[10] For other statements of the dynamical systems perspective, see [4, 21, 27].

group activity of a number of real neurons. But it is implausible to postulate a general explanation of situated activity on the basis of recurring correspondences between occurrent point attractors in neural activation space and meaningful elements of the environment. It seems likely that real neural systems do not exhibit activation space dynamics which succeed in converging to point attractors — except at the point of death [25].

At first sight, this is not enough to threaten representational forms of explanation in the study of situated activity, because surely, in principle, occurrent periodic attractors in a neural control system could be detected and used by other processes in that control system; and, despite the famous 'sensitive dependence on initial conditions,' a chaotic attractor is still bounded and, therefore, as detectable as its cyclic cousin. So, in theory, at least some periodic or chaotic attractors in the activation space of a dynamical neural network could assume representational status through (i) being causally correlated with states of the environment and specific adaptive behaviours, and (ii) being used by identifiable sub-systems in the control architecture. In theory, maybe, but it is overwhelmingly likely that dynamical neural networks, coupled to uncertain, changing environments, will consistently fail to reach *any* stable attractors in activation space, even though the networks may spend periods on transients which are in the basins of attraction of such stable dynamical states. For example, Yamauchi and Beer [29] show that evolved dynamical neural networks are capable of successfully performing relatively abstract tasks involving sequential behaviour, such as generating a fixed sequence of outputs in response to external decision triggers. One general feature of the various successful network-dynamics is that, as long as environmental triggers are forthcoming, the networks do not necessarily reach the attractors which arise in activation space in response to the triggering stimuli. The networks spend most of their time producing the desired behaviour whilst on the transients of those attractors.

Any realist about representation must expect the putative internal representational states to play a particular kind of role in the production of behaviour — a causal role which is interpretable according to the principles of homuncularity (as identified in section 3). Theoretically reidentifiable (but, in practice, largely non-occurring) attractors cannot play that role. It may seem that the requirement is 'merely' for some (non-magical) story about how the behaviour of the sub-systems of the overall control system can be causally affected by transient-dynamics. But even if such 'practical' details were worked out, it is still the case that causal correlations between transients and environmental states are not, on their own, sufficient to reestablish representationalism. The salvage operation makes sense only if we are talking about a control system for which the conceptual framework of homuncular decomposition is appropriate or useful. The causal interactions between identified sub-modules need to occur in a such a way that it makes sense to speak of the detection and use of information by homunculi. But if the interacting sub-modules identified by our analysis of the nervous system are themselves more properly viewed as coupled dynamical systems, then, in general, it will not be useful to interpret the causal interactions between those modules as representation-passing communications between hierarchically-organized homunculi.

It is worth driving that last point home. Mainstream connectionist networks are generally understood as pattern completion devices. Indeed their famous achievements in associative memory, learning, default reasoning and flexible generalization are the results of this more basic capacity for pattern completion. It is by virtue of imposing the sorts of restrictions on the dynamics of artificial neural networks discussed in section 5 that networks can be used as (for example) associative memory devices [4]. We have already seen that if the restrictive architectural assumptions are relaxed, then more complex dynamical behaviour ensues. But now once we have entered the realm of dynamical neural

networks coupled to one another and to changing, uncertain environments, it is surely idealistic in the extreme to suppose that the trajectories of such networks will be so constrained, that the description of network-dynamics as a process of pattern completion will remain accurate. This is relevant to the applicability of homuncular decomposition; a pattern-completing network, with its well-defined and well-behaved input-output profile, is a highly suitable applicant for the job of sub-personal cognitive homunculus. By contrast, a coupled dynamical neural network would be unlikely to get an interview! But, if we should not expect adaptive control systems to respect the principles of homuncularity (see section 3), then the fact that dynamical networks may be unsuitable as homunculi does not tell against their status as control systems for situated agents.

What about computation? Well, of course, if computation is defined as the manipulation of (or transformation between) representations, then doubts about representation undermine the computational half of the double-act. But it is worth noting that there are semi-independent reasons for questioning the suitability of computational explanation. According to the orthodox approach, the mathematics appropriate for describing situated agency are the mathematics of computation, as described in the theories of logic and computability theory [20]. In this strict sense of 'computation,' the processing — the manipulation of representations — is discrete, and the processing cycles can be divided into temporally distinct stages of input, computation and output. By contrast, I have suggested that the general behaviour of a dynamical neural network control system, interacting with other dynamical networks and a dynamic environment, will not be interpretable as a series of discrete state transitions realizing isolable input-output computations. Thus the dynamical systems perspective rejects the mathematics of computation in favour of the fundamentally continuous mathematics of dynamical systems theory, and replaces the input-output transition cycle with the richly interactive dynamics of coupling.

So far, I have concentrated on 'negative' arguments which question the explanatory primacy of representations and computations, with only an occasional reference to positive examples of dynamical systems explanations. But I shall close this section with an all-too-brief description of work by Husbands, Harvey and Cliff in which concepts and tools from dynamical systems theory are crucial in providing an explanation of the situated behaviour of a simulated robot. (For full details, see [21].)

A visually-guided robot is placed in a circular arena with a black wall and a white floor and ceiling. From any randomly chosen position in the arena, the robot's task is to reach the centre as soon as possible and then to stay there. To this end, control systems are developed through the artificial concurrent evolution of dynamical neural networks and sensor-morphologies (as per the evolutionary robotics methodology described in section 6). Despite the simplicity of the scenario, the results are suggestive. For example, consider one controller-network — called 'C2' — which was evolved in an environment with a wall-height of 15. The first stage of the analysis was to understand the dynamics of the evolved network. The channels of activation in the network flow in complex and counter-intuitive ways, due to the complex nature of the connectivity and the existence of feedback loops. By eliminating redundant units and connections (which may be left over from earlier evolutionary stages) the significant visuo-motor pathways are identified, as are the conditions under which those pathways become active. The second stage is to combine the network analysis with a two-dimensional state space representing (to the observer) the robot's visual world (i.e., the visual signals which would be received at different positions in the world). The result is a phase portrait predicting the way in which — according to the dynamical systems model of the 'physical' system — the robot will tend to move through the state space. The phase portrait features a single point attractor in visuo-

motor space (not activation space!) corresponding to a very low radius circle about the centre of the world, and the whole state space is, in effect, a basin of attraction for this attractor. In short, the model predicts that the robot will *always* succeed at its task, a prediction which was borne out by empirical demonstration.[11]

The next stage was to investigate the adaptiveness of the control system by analysing the behaviour of the robot in an arena with wall-height 5, i.e., in an environment for which the control dynamics were not specifically evolved. The change in wall-height means a change in the structure of the robot's visual state space. The same process of analysis now yields a phase portrait featuring two point attractors in visuo-motor space, both corresponding to successful behaviours. Once again the model was confirmed by empirical demonstration. So the dynamical systems analysis correctly predicts that the control system is general in that it will achieve its goal by exhibiting different situated dynamics in different environments. The moral of this example is that in a particular instance of situated activity, certain aspects of the activation space dynamics of the controlling network may play a crucial part in our explanation, because what we will need to understand is how the dynamics of the control system interact with the dynamics of the agent's environment to produce well-tuned adaptive behaviour. Internal dynamics alone — representational, computational, or otherwise — will not be sufficient to gain a general understanding of embedded cognition, because, in most cases, amputating the agent's environment from the explanation will leave one with no explanation at all. Thus the dynamical systems approach develops a key insight which was present in behaviour-based robotics: not only is it a good idea to use the environment as a source of information, but any adequate explanation of adaptive situated activity will have to include (in a fundamental way) the environment of the creature under investigation.

8 Conclusions

I have argued that the explanatory frameworks adopted in classical A.I. and orthodox connectionism should no longer be assumed in the general study of situated agency. But old dogs can (sometimes) learn new tricks. Nothing in this paper rules out the possibility that new uses for the terms 'representation' and 'computation' will be found. Indeed, the notion of deictic representation employed in behaviour-based robotics is an example of one such revision. Perhaps we can force 'representation' and 'computation' to be the key theoretical terms in a science of situated activity, if we fiddle with their orthodox interpretations. But, in a cantankerous frame of mind, I am compelled to ask "why bother?" One of the benefits of the orthodox notions of representation and computation is the fact that they are well-defined enough to permit empirical investigations of their applicability in any particular case of adaptive behaviour. We should be loathe to discard the undeniable advantages of this property. The explanatory territory of orthodox cognitive science may be precisely the sort of reasoning that does *not* involve non-arbitrary sensory-motor coupling with an environment. But then how much cognition will this account for in the non-human regions of the animal kingdom? And how much *human* cognition is actually 'disembodied' and 'non-situated' in this way? Given the 'simple-systems-first,' incremental approach so sensibly adopted in A-Life, I am compelled to fall back on that old chestnut of empirical work, 'only time will tell.' But, somewhere in the distance, I can hear a clanging noise —

[11] It is important to stress that certain details of this particular dynamical analysis are contingent upon the nature of the specific scenario (e.g., the symmetries of the environment which allow the observer to generate the state space). It is the *style* of analysis which is at issue.

and it's getting louder.[12]

References

[1] R. H. Abraham and C. D. Shaw. *Dynamics - The Geometry of Behaviour 2nd edition.* Addison-Wesley, Redwood City, California, 1992.

[2] M. A. Arbib. *The Metaphorical Brain 2: Neural Networks and Beyond.* John Wiley and Sons, New York, 1989.

[3] R. Beer. *Intelligence as Adaptive Behaviour: An Experiment in Computational Neuroethology.* Academic Press, San Diego, California, 1990.

[4] R. Beer. A dynamical systems perspective on autonomous agents. Technical Report 92-11, Case Western Reserve University, Cleveland, Ohio, 1992.

[5] R.D. Beer and J.C. Gallagher. Evolving dynamic neural networks for adaptive behaviour. *Adaptive Behavior*, 1:91–122, 1992.

[6] M. A. Boden. Autonomy and artificiality. To be published in *A.I.S.B. Quarterly* and D. Cliff (ed.) *Evolutionary Robotics and Artificial Life* (provisional title), forthcoming.

[7] R. Brooks. Intelligence without reason. In *Proceedings of the Twelth International Joint Conference on Artificial Intelligence*, pages 569–95, San Mateo, California, 1991. Morgan Kauffman.

[8] R. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–59, 1991.

[9] P. S. Churchland. *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* MIT Press / Bradford Books, Cambridge, Massachusetts, 1986.

[10] A. Clark. *Associative Engines: Connectionism, Concepts, and Representational Change.* M.I.T. Press / Bradford Books, Cambridge, Massachusetts and London, England, 1993.

[11] D. Cliff. Computational neuroethology: a provisional manifesto. In J.-A. Meyer and S. W. Wilson, editors, *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behaviour*, Cambridge, Massachusetts, 1991. M.I.T. Press / Bradford Books.

[12] D. Cliff, I. Harvey, and P. Husbands. Explorations in evolutionary robotics. *Adaptive Behavior*, 2:73–110, 1993.

[13] D. Cliff, P. Husbands, and I. Harvey. Analysis of evolved sensory motor controllers. In *Proceedings of the Second European Conference on Artificial Life*, pages 192–204, 1993.

[14] J. H. Connell. A colony architecture for an artificial creature. A.i. memo 1151, M.I.T. A.I. Lab, 1989.

[12] This work was supported by British Academy major award no. 92/1693. Many thanks to Maggie Boden, Dave Cliff and Matthew Elton for their valuable comments on an earlier version of this article.

[15] N. Franceschini, J-M. Pichon, and C. Blanes. Real time visuomotor control: from flies to robots. In *Proceedings of: International Conference on Advanced Robotics, Pisa*, 1991.

[16] M. Giunti. *Computers, Dynamical Systems, Phenomena, and the Mind.* PhD thesis, Department of History and Philosophy of Science, Indiana University, 1991.

[17] I. Harvey. Untimed and misrepresented: Connectionism and the computer metaphor. Cognitive Science Research Paper 245, University of Sussex, 1992.

[18] I. Harvey, P. Husbands, and D. Cliff. Issues in evolutionary robotics. In J. A. Meyer, H. L. Roitblat, and S.W. Wilson, editors, *From Animals to Animats 2: Proceedings of the Second International Conference on the Simulation of Adaptive Behavior*, pages 364–73, Cambridge Massachusetts and London, England, 1993. MIT Press / Bradford Books.

[19] I. Harvey, P. Husbands, and D. Cliff. Seeing the light: Artificial evolution, real vision. In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, forthcoming.

[20] T. Horgan and J. Tienson. A nonclassical framework for cognitive science. *Synthese — special issue on Connectionism and the Frontiers of Artificial Intelligence*, forthcoming.

[21] P. Husbands, I. Harvey, and D. Cliff. Circle in the round: State space attractors for evolved sighted robots. *Proceedings of the N.A.T.O. Advanced Study Institute Workshop on the Biology and Technology of Intelligent Autonomous Agents*, forthcoming.

[22] H. R. Maturana. Biology of cognition. In *Autopoiesis and Cognition: the Realization of the Living.* Reidel, Dordrecht, 1970. Published in 1980, the book contains the piece by Maturana (dating from 1970) and a piece by F. Varela.

[23] A. Newell and H. Simon. Computer science as empirical enquiry. In J. Haugeland, editor, *Mind Design.* M.I.T. Press, Cambridge, Massachusetts, 1976.

[24] N. J. Nilsson, editor. *Shakey the Robot.* S.R.I. A.I. Centre, April 1984. Tech. Rept. no.323.

[25] C. A. Skarda and W. J. Freeman. How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, 10:161–195, 1987.

[26] P. Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74, 1988.

[27] T. J. van Gelder. What might cognition be if not computation? Technical Report 75, Indiana University, Cognitive Sciences, 1992.

[28] S. W. Wilson. Knowledge growth in an artificial animal. In J. J. Grefenstette, editor, *Proceedings of an International Conference on Genetic Algorithms and their Applications*, pages 16–23, Pittsburg, PA and Hillsdale, New Jersey, 1985. Lawrence Erlbaum Associates.

[29] B. M. Yamauchi and R. D. Beer. Sequential behavior and learning in evolved dynamical neural networks. Available as Case Western Reserve University Tech. Report no. CES-93-25. Cleveland, Ohio, Submitted to *Adaptive Behavior.*