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.

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

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.

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,

5 hinking about Networks

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].

"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.

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

group activity of a number of real neurons. But it is implausible to postulate a general ex-

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 beould92f

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 sys-

and it's getting louder.[12]

References

[1] R. H. Abraham and C. D. Shaw. *Dynamics - The Geometry of Behaviour 2nd edition.*

[15] N. Franceschini, J-M. Pichon, and C. Blanes. Real time visuomotor control: from flies