The Dynamics of Associative Learning in an Evolved Situated Agent

Eduardo Izquierdo and Inman Harvey

Centre for Computational Neuroscience and Robotics
Department of Informatics, University of Sussex, Brighton, UK
{e.j.izquierdo, inmanh}@sussex.ac.uk

Abstract. Artificial agents controlled by dynamic recurrent node networks with fixed weights are evolved to search for food and associate it with one of two different temperatures depending on experience. The task requires either instrumental or classical conditioned responses to be learned. The paper extends previous work in this area by requiring that a situated agent be capable of re-learning during its lifetime. We analyse the best-evolved agent's behaviour and explain in some depth how it arises from the dynamics of the coupled agent-environment system.

1 Introduction

Learning is a behaviour. In fact, it is a *change* of behaviour over time. Living organisms show a variety of behaviours that are modulated by environmental conditions and previous experience. A major goal of the artificial life sciences is to elucidate the dynamical bases of such experience-dependent adaptive behaviour.

Associative learning is a particularly adaptive form of such experience modulated behaviour, as it requires responses to be paired with a particular stimulus. Organisms at several levels of 'complexity' provide evidence for this, including many extraordinarily simple ones. In the small nematode worm *C. elegans*, evidence for the formation of associations between temperatures and food has been known for quite some time [5]. However, the mechanisms required for the storage and resetting of this memory are still largely unknown.

In the animal learning theory there is the idea of the strengthening of a 'connection' between a stimulus and a response. This has been directly translated to the strengthening of physical connections between neurons. While this is a good description at the level of the agent's interaction with the environment (behavioural description), there need not be a direct correspondence of connection-forming processes in the internal behaviour-producing mechanisms of the agent. We believe there is a more fundamental principle underlying learning behaviour at the level of an organism's internal mechanisms that has to do with dynamics on multiple timescales.

The aim of this work is to: (1) successfully evolve the smallest possible integrated dynamical system controller with fixed weights in a situated[1] agent on an

[1] By situated we mean an agent that is embedded in a world; and thus its ongoing sensori stimuli is dynamically determined by its own actions.

associative learning task requiring re-learning, (2) perform a behavioural analysis of the best evolved agent; and (3) study the coupled agent-environment dynamics of a successful controller and attempt to understand it as implementing a finite state machine (FSM), so as to compare with similar work [6].

2 Related Work

A number of researchers have used genetic algorithms to evolve, for tasks requiring associative learning, dynamical neural controllers without in-built synaptic plasticity mechanisms. Yamauchi and Beer [8] were the first to explore this idea using a one-dimensional navigation task with a goal and a landmark. Attempts to evolve an 'integrated' network failed, so a modular approach was taken.

Blynel and Floreano [2] evolve controllers on a relatively similar task and environment. In their version, because the light is fixed to one side of the arena and the goal is the only thing that changes, it is possible for the agent to employ a reactive turn left or right strategy, as opposed to approaching or avoiding the light; making it unnecessary to form an association between light and goal.

Attempts to remedy those initial difficulties were successfully overcome by Tuci et al. [7] in a two-dimensional version of the same task. As the emphasis of that work was on the evolutionary process, no further analysis of the behaviour or internal dynamics was performed.

Fernando in [3] explores the same associative learning task in a slightly more complicated T-maze environment. Despite not being able to evolve an agent that solves the task completely, an analysis of the best performing agent in terms of animal learning theory is attempted.

Such work demonstrates that multitimescale dynamics can exhibit learning-like behaviour without synaptic plasticity mechanisms. However, none of the previous work deals with re-learning during the lifetime of the agent: the agent's internal state is reset when tested on a different environment. Also, the internal mechanisms of the best-evolved agents have not been explored in much depth or at all in some cases.

Phattanarasri et al. [6] study in-depth the dynamics of an evolved circuit for an associative learning task very similar to the one being presented here. The main difference with this work is that their experiments take place in a non-situated agent. Of particular interest is their analysis of the evolved internal mechanisms, which can be understood to implement a FSM.

3 Methodology

We use evolution to synthesize continuous-time recurrent neural networks that display associative learning behaviour when situated. The task is loosely abstracted from the temperature preference behaviour observed in the nematode worm *C. elegans* [5]. In particular, we would like an agent that is capable of associating temperature with food in two different types of environment, and re-learning: modify its temperature preference during its lifetime when required.

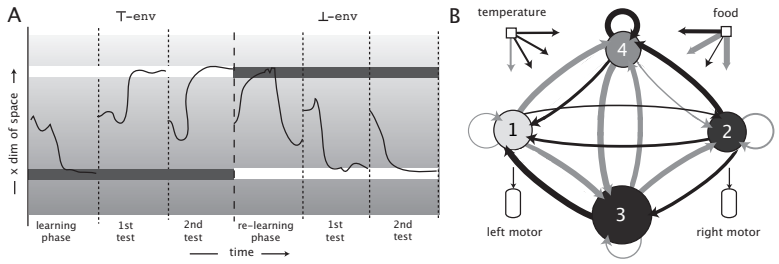

Fig. 1. [**A**] Example trial. 1D projection of environment with thermal gradient (shades of grey). 'Nutritious food' denoted by white bars; 'poisonous' with black. [**B**] Agent architecture with 4 fully inter-connected nodes, a food and a temperature sensor, and two wheels controlled by arbitrarily chosen nodes. Parameters of the best evolved circuit are also depicted. Nodes are shaded according to their bias. Excitatory connections (black) and inhibitory (grey), with the width of the line proportional to the strength. Time-constants represented by size, with larger circles representing slower nodes.

We use a 2D arena with a thermal gradient along one of its dimensions containing two types of food: 'nutritious' and 'poisonous'. Each type of food can be found only in regions in a particular temperature range: '*hot*' between [9,10]; '*cold*' between [-10,-9]. Which region the nutritious food can be found in depends on the type of environment: ⊤-env, nutritious food in the *hot* region; and ⊥-env, in the *cold* region. For each of the different environment types, the poisonous food can be found in the opposite region to the nutritious food. There are no walls and the thermal gradient extends in all directions.

An example trial of the task is depicted in Figure 1A. The task involves placing an agent at random in the central region (between [-2,2]) of the arena (including random orientation) in one of the two environment types, requiring it to find and stay on the food as efficiently as possible. The first challenge involves exploring the whole of the arena in search for food. After a random amount of time (between [80,100] units), the agent is physically displaced back towards the central region of the arena and given a random orientation again. A successful agent should navigate up or down the thermal gradient depending on whether it had found food in the *hot* or *cold* region in the previous trial, respectively. This requires that it learn and remember in which of two environment types it finds itself. Less frequently, the displacement involves changing the environment type as well. This requires that the agent remain sufficiently plastic to change its temperature preference online. Although it is this learning and re-learning phenomena that are central to our paper, there is also a more basic sensory-motor challenge involved in navigating up and down the thermal gradient which will not be explored.

Agents are modelled as circular bodies of radius 1 with two diametrically opposed motors and two sensors. Agents can move forwards and turn. The mass of the body is sufficiently small so that the motor's output is the tangential velocity at the point where the motor is located. The agent can sense the local

temperature in the environment as well as the food. The food, however, cannot be perceived unless the agent is directly upon it. The food sensor is: 1 for nutritious food, -1 for poisonous food, and 0 when no food is present. The temperature sensor can have any real value.

For the internal dynamics of the agent, we use a continuous-time recurrent neural network (CTRNN) with the following state equation [1]:

$$\tau_i \dot{y}_i = -y_i + \sum_{j=1}^{N} w_{ji}\sigma\left(y_j + \theta_j\right) + s_i T(x) + g_i F(x;e) \tag{1}$$

where y is the activation of each node; τ is the time constant; w_{ji} is the strength of the connection from the j^{th} to the i^{th} node; θ is a bias term; $\sigma(z) = 1/(1+e^{-z})$ is the standard logistic activation function; $T(x)$ is the thermal sensor, a function of the agent's position along one of the dimensions of the physical space, x; s_i is the strength of the connection from the thermal sensor; $F(x;e)$ is the food sensor, also a function of x but parameterized by the type of environment, e; g_i is the strength of the connection from the food sensor; and N represents the number of nodes in the network. In simulation, node activations are calculated forward through time by straightforward time-slicing using Euler integration with a time-step of 0.1. The network is fully connected (see Figure 1B). There are no additional weight changing or any other parameter changing rules.

The connection weights, biases, and time-constants in Equation 1 are encoded in a genotype as a vector of real numbers and evolved using the microbial genetic algorithm [4]. The size of the population used was 50. We define a generation as the time it takes to generate 50 new individuals.

The fitness of a circuit is obtained by minimising the relative distance away from the food at the beginning of each test (a), and maximising the time spent sensing food towards the end of the same phase (b), according to

$$a = \int_{t=0}^{50} \left(\frac{20-d}{20}\right) dt \ , \quad b = \int_{t=30}^{80} F dt \tag{2}$$

where F is the agent's sensor for food and d is the absolute distance between the source of food and the position of the agent capped at 20. Both components are normalised to run between 0 and 1. The two components are clearly linked: the first provides emphasis on heading in the direction towards where the food should be at the start of the trial; the second emphasizes staying directly on top of the food once found.

A fitness trial consists of the evaluation of an agent's performance for the number of times it is displaced in the same environment type, p, and for the number of changes of environment type, k, all without reinitialising the agent's state. No evaluation takes place at the start of a trial, nor immediately after a change of environment type. This is repeated 50 times for each individual and the fitness taken from the multiplication of their averages, $f = \bar{a} \cdot \bar{b}$. Each repetition involves the re-initialisation of the agent's internal state.

Following [6], a set of evolutionary stages of increasing complexity are employed. The changes are in the starting orientation of the agent, φ, after each

start of trial or displacement; in the number of times an agent is tested (i.e. displaced), k; and the number of changes of environment type, p; as follows:

Stage	1	2	3	4	5
φ	$\{0, \pi\}$	$[0,2\pi)$	$[0,2\pi)$	$[0,2\pi)$	$[0,2\pi)$
k	1	1	1	2	5
p	1	1	5	[1,5]	[1,5]

Transitions occur when the best fitness exceeds 0.8 consistently (i.e. for 5 consecutive generations). At the last stage, the orientation is chosen at random from the full range, the environment type changes 5 times during the agent's lifetime, and the changes occur between the first and the fifth displacement at random.

4 Results

4.1 Evolutionary performance

We attempted evolving 3, 4, and 5-node circuits for this task using 15 evolutionary runs with different seeds for 10000 generations each. The proportion of evolutionary runs that reached the different stages are depicted in Figure 2A. While no 3-node populations reached the last stage, several 4 and 5-node populations did. In fact, the majority of 5-node runs were highly successful, but we will focus our attention on the smallest successful circuit obtained. The interest in evolving the smallest circuit that solves the task is primarily to make the analysis most amenable to the mathematics of dynamical systems theory.

An example evolutionary trajectory for the population that produced the best 4-node agent is shown in Figure 2B. As can be seen, the fitness drops sharply after every transition except the last: once the circuit is able to generalize to all learning scenarios. It is the best agent of this evolutionary run that will be analysed in some depth in the rest of this paper.

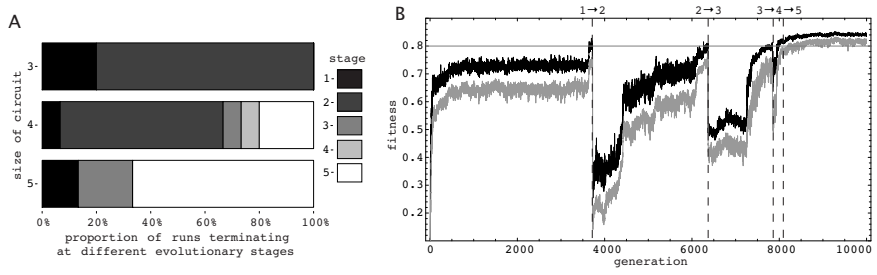

Fig. 2. Evolutionary performance. [**A**] Proportion of populations that terminated at a certain evolutionary stage for different size circuits. [**B**] Fitness vs. generation for the best evolved 4-node population (best in black and average in grey). Transitions between stages (dashed lines) occur when the best fitness consistently exceeds the horizontal grey line and are labelled accordingly.

4.2 Behavioural analysis

The performance of the best circuit was further tested using 10^4 evaluation trials, each with 10 changes of environment type, between [1,10] displacements, noise in the sensors and motors drawn from a Gaussian distribution (σ=0.05), and a time-step an order of magnitude smaller (0.01). As we are interested in how well the agent finds the nutritious food in the face of changing environments, only the b component of fitness is considered. The best 4-node circuit obtained 98.81% success rate on this test, meaning that it generalises well on a broad range of situations. Since the slope of the thermal gradient remains constant throughout evolution, the agent could use the distance instead of the temperature as the relevant factor to remember. We used the same test while varying the slope of the gradient between $\pm$20% with the success rate dropping by only a minor fraction (98.48% success), meaning the agent relies on the temperature and not the distance the food is away from the centre.

Figure 3 shows the behaviour of this agent on a typical sequence trial with 2 changes of environment type. At the beginning of the trial, the agent navigates down the thermal gradient but switches to navigating up before reaching the usual region where food could have been located. This is part of the search strategy, as it does not yet know in what type of environment it finds itself. When displaced for the first and second times after reaching the food, however, it navigates more directly up the thermal gradient. Subsequently the environment type is changed, unaware the agent navigates up the thermal gradient as for previous trials, with the difference that negative reinforcement is encountered (but only very briefly[2]). The agent navigates past this food region and eventually changes behaviour to navigate in the opposite direction of the gradient, until reaching the nutritious food on the *cold* region. On subsequent trials, the agent will navigate directly down the gradient, showing that it remembers where the food was last found in the other type of environments as well. A similar pattern is observed in the second change of environment type. This demonstrates the agent's ability to learn and remember its past behaviour, as well as the flexibility to remain plastic to ongoing changes in the environment type.

We note that all 4 nodes are active at one point or another during the sequence trial; with most of the activity occurring during the navigation phase. Particularly interesting is the activity of node o_3, which seems to be the only one keeping track of which environment type it finds itself in. This is also the node with the largest time constant in the circuit; all other nodes are as fast acting as allowed (see Figure 1B).

Before any experience, does the agent navigate up or down the thermal gradient? and what does this depend on? We studied the long-term behaviour of the agent when initialised in an environment with nutritious food on both *cold* and *hot* regions. As can be seen in Figure 4A, what the agent does depends mainly on its starting position: visiting the furthest region first. How does experience affect this pattern? After learning has occurred, the agent will preferably head

[2] Absence of poisonous food in the environment does not affect the learning behaviour in this agent. The reason is that the negative reinforcement is redundant in this task.

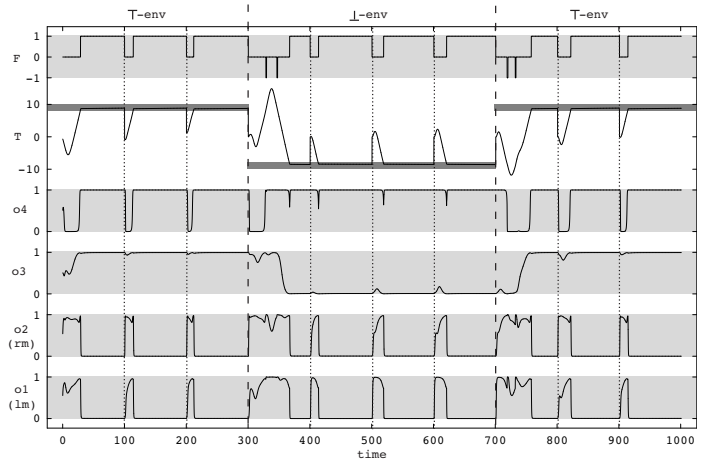

Fig. 3. Activity of the best 4-node circuit on a typical trial sequence. From top to bottom the traces correspond to the food signal (F), the temperature signal (T), and the outputs of the neurons (o_i). The last two neurons control the right (rm) and left (lm) motors. The dark grey horizontal bars in the temperature trace depict where nutritious food is to be found for that trial ($\top$ or $\bot$). Dotted vertical lines mark different trials (where the agent is displaced). Dashed lines mark transitions between environments.

towards *hot* or *cold* regions, even with nutritious food on both, depending on where food was found in the previous trial (see Figures 4B and 4C, respectively). This shows how behaviour is appropriately modulated according to previous experiences regardless of initial position and orientation.

4.3 Dynamics of the coupled agent-environment system

We next turn to the dynamics underlying the behavioural phenomena described in the previous section. The primary interest is in understanding how this agent's dynamics is structured so that where food was encountered in the past affects which direction of the thermal gradient it will navigate towards. From the equations describing the coupled agent-environment system we can make some general observations. First, the agent is a nonautonomous dynamical system with two inputs, T and F. Second, although T varies continuously as a function of x, discontinuities are introduced into the dynamics by the food sensor because $F(x;e)$ is a discontinuous function of x, making the agent a hybrid dynamical system. Given these two factors, the best way to study its operation is to characterize its autonomous dynamics for all possible combinations and then examine the transient dynamics induced by the agent-environment interaction. If we take into consideration only the range of temperatures where the agent was observed to navigate around, then there are five possible bifurcation diagrams to consider: $P_{\pm 15}$ (temperature between [-15,15] with no reinforcement), $P_{\downarrow +}$ (*cold* temp. and positive reinf.), $P_{\uparrow +}$ (*hot* temp. and positive reinf.), $P_{\downarrow -}$ (*cold* temp.

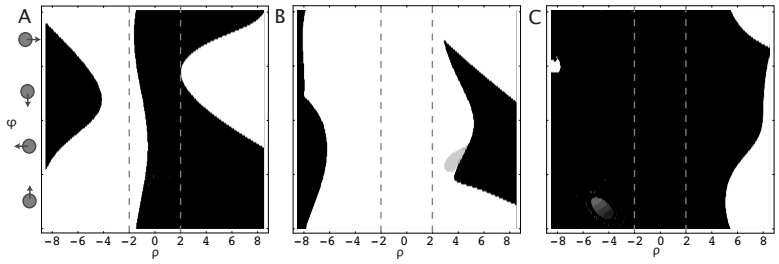

Fig. 4. Points in the map represent the average position (over 20 repetitions) of the agent after 100 units of time with nutritious food on both *hot* (white) and *cold* (black) regions while varying its starting position (p) and orientation (φ). Points in-between are in shades of grey. Grey dashed lines mark the conditions for which the agent was evolved. Different maps show the agent's behaviour with different past experiences: [**A**] Before any experience. [**B**] After $\top$-environment. [**C**] After $\bot$-environment.

and negative reinf.), and $P_{\uparrow -}$ (*hot* temp. and negative reinf.). Three-dimensional projections of the stable solutions of the first three of these are shown in Figure 5A, coded in shades of grey as a function of the temperature and labelled accordingly. The portraits corresponding to the negative reinforcements can be left out of the analysis because they do not affect the performance of the agent's learning behaviour. As can be seen, for mid-temperatures ($P_{\pm 9}$) the long-term behaviour of the system is bistable. As the temperature increases or decreases outside of this range, only one attractor is left in opposite ends of the original for *cold* and *hot*. Similarly, for $P_{\downarrow +}$ the dynamics are bistable and for $P_{\uparrow +}$ there is only one stable state.

How do these bifurcation diagrams combine to produce the learning behaviour? We can study the transient trajectories in the internal state of the agent as it interacts with its environment. In Figure 5B we show a set of trajectories from behaviours crucial for the task using the same projection as in the previous plot. Can we interpret the transitions in the internal state of the agent as implementing a FSM? We were unable to do so. The difficulty arises from the agent's dependence on the temperature sensing as an ongoing and continuous perturbation. We hypothesize that it is the discretisation and non-situatedness of the task in [6] that facilitates their FSM interpretation. Only when we consider a different form of state machine that allows for ongoing sensori-motor interactions can we summarize the coupled agent-environment system in relation to the agent's internal dynamics. We will call this an 'interactive state machine' (see Figure 5C). Although similar, strictly speaking the diagram is not a FSM because some of the states include ongoing interactions with the environment. In it, the finite states the system can be in are denoted by circles labelled: $\uparrow +$ or $\downarrow +$, for when nutritious food is found in the *hot* and *cold* regions, respectively. The graded ellipses represent the 'interactive states': where the agent's state moves it in relation to the environment, and the change of temperature changes the dynamics of the agent in turn. There are two of these: $\uparrow$ and $\downarrow$, for what results

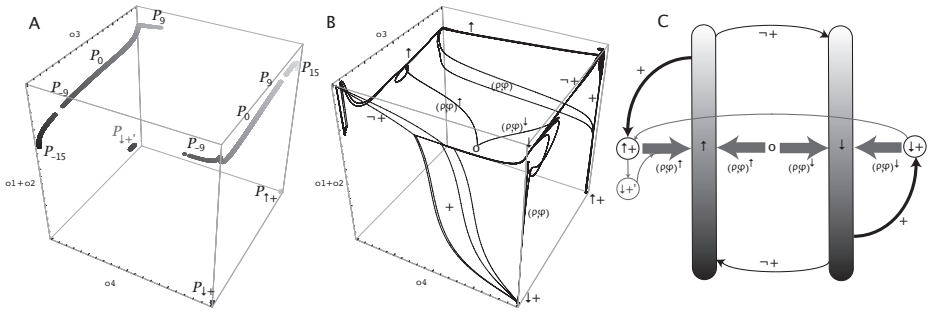

Fig. 5. Agent-environment coupled dynamics. [**A**] Equilibrium points of the nonautonomous system depending on temperature (shade of grey) and positive reinforcement. [**B**] 3D projection of the trajectories in internal space state for a typical set of behaviours. See main text for the labels. [**C**] Diagram of the coupled dynamics.

in navigation up or down the gradient, respectively. We denote the starting internal state as o. Physical displacements events are depicted with thick arrows. We can characterise the basins of attraction of the bistable dynamics in $P_{\pm 9}$ as a function of the agent's position and orientation from Figure 4A as $(p, \varphi)^z$, where z represents the long-term behaviour ($\uparrow$ or $\downarrow$). The black arrows denote the encountering of nutritious food, $+$. The thin arrows connecting the ellipses denote the transition from one stable state in $P_{\pm 15}$ to the other in the internal dynamics. This occurs when the agent reaches colder or hotter temperatures. The diagram up to this point is sufficient to fully characterise the observed behavioural phenomena. There is an additional finite state that is never reached during regular associative learning which we denote as $\downarrow+$'.

4.4 Predictions from the dynamics

The study of the dynamics suggest a number of predictions which we could confirm using behavioural studies. Although a full study of the predictions would require further space, two of them are mentioned briefly. First, as a result from the bistability of $P_{\downarrow+}$, we could predict and confirm that even after experiencing environments with food in the *cold* regions, if exposed to *hot* temperatures and food simultaneously for sufficiently long, the agent could be re-conditioned to navigate up the thermal gradient. This was not the case in the opposite scenario, where the agent required doing the down-the-thermal-gradient navigation behaviour to remember. We can describe the agent as employing a mixture of classical (pairing two signals) and operant (pairing an action with a reinforcement) conditioning. Second, and as a consequence of the geometry of $P_{\pm 15}$, we could predict and confirm that in the total absence of any kind of food, the coupled system falls into a limit cycle, that involves the agent switching between going up and down the gradient modalities. Although this was not a scenario the agent was evolved for, it could be interpreted as a higher level 'searching for food' behaviour that emerges from the lower level behaviours selected for.

5 Concluding Remarks

We successfully evolved *situated* agents with fixed weight dynamical neural controllers on an associative learning task requiring *re-learning*. The observed phenomena can be described as the ability to perform two different behaviours and appropriately switch between them when necessary using feedback from the interactions with the environment. The question of whether such experience-dependent behaviour is actually 'learning' is discussed in more depth in [6]. The dynamics of the coupled agent-environment is explored in some depth. Attempts to generate a FSM are unsuccessful but a form of 'interactive state machine' is provided instead. From the dynamics, two predictions are explored.

This work raises a number of issues we believe deserve to be further studied. First, in the case of a situated agent, how useful is the conventional distinction drawn between operant and classical conditioning? Our work suggests that the distinction arises from the discretisation of the task or the minimisation of the coupling between agent and environment. Second, in such 'representationally-hungry' tasks, correlations between the activity of internal components and that which the agent has to remember are trivial to spot. Could they be interpreted as symbols the agent can manipulate to perform computations? Further work unravelling what is meant by 'internal representations' from minimal model systems such as the one presented here should be of interest. Finally, an important next step will be to extend this work to an agent that can associate any temperature along a continuum with food, as is the case in the phenomena observed in *C. elegans* from which this task was abstracted.

References

1. R.D. Beer. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509, 1995.
2. J. Blynel and D. Floreano. Levels of dynamics and adaptive behavior in evolutionary neural controllers. In *Proc. of the 7th Int. Conf. on Simulation of Adaptive Behavior: From animals to animats*, pages 272–281. MIT Press, 2002.
3. C. Fernando. A situated and embodied model of classical and instrumental learning. Master's thesis, COGS, University of Sussex, 2002.
4. I. Harvey. Artificial evolution: a continuing SAGA. In T. Gomi, editor, *Evolutionary Robotics: From Intelligent Robots to Artificial Life*. Springer-Verlag, 2001.
5. E.M. Hedgecock and R.L. Russell. Normal and mutant thermotaxis in the nematode caenorhabditis elegans. *Proc. Nat. Acad. Sci. USA*, 72(10):4061–4065, 1975.
6. P. Phattanasri, H.J. Chiel, and R.D. Beer. The dynamics of associative learning in evolved model circuits. *Adaptive Behavior*, Submitted.
7. E. Tuci, M. Quinn, and I. Harvey. An evolutionary ecological approach to evolving learning behavior using a robot based model. *Adaptive Behavior*, 10(3/4):201–221, 2003.
8. B.M. Yamauchi and R.D. Beer. Integrating reactive, sequential and learning behavior using dynamical neural networks. In D. Cliff, P. Husbands, J. Meyer, and S. Wilson, editors, *From Animals to Animats 3: Proc. of the Third Int. Conf. on Simulation of Adaptive Behavior*, pages 382–391. MIT Press, 1994.