The Role of Behavioral Extinction in Animat Action Selection

Mark Witkowski
Department of Electrical and Electronic Engineering
Imperial College of Science, Technology and Medicine
Exhibition Road
London SW7 2BT
United Kingdom
m.witkowski@ic.ac.uk

Abstract

Behavioral Extinction is a long established experimental procedure in animal learning studies that discover what happens when Action Selection fails. This paper integrates the results of these studies into the existing Dynamic Expectancy Model, and considers the impact this has on the Action Selection and learning properties of that model. A series of experimental investigations is presented to illustrate how the behavioral extinction mechanism can be combined with existing properties of the model to protect the Animat from harm in circumstances where Action Selection would otherwise fail.

1 Introduction

Action Selection has emerged recently as a key issue in the modelling of biological systems. Put simply, it asks the question "given the current circumstances, and taking everything into account, which action or behavior pattern should an animal (or Animat) select to do right now." The answer, it seems, is far from straightforward, but four factors are consistently identified as strongly implicated:

- The current context, as identified by the animal's sensory and proprio-receptive apparatus
- Past, and in particular, recent past, experiences of the animal (learning)
- Goals or motivating requirements, internal to the animal
- The innate, or pre-programmed, capabilities of the animal

This paper will consider what happens when an Animat model is taken outside its normal envelope of Action Selection operation, and investigates what an Animat might do under increasingly adverse conditions. The paper notes detailed findings from laboratory studies in animal learning, in particular that of the *Behavioral Extinction* phenomena. It integrates these findings into an established Animat model and then considers the consequences these results might have for Animat systems in the broader context of Action Selection. Most natural learning phenomena are considered reversible to a greater or lesser extent, and extinction studies investigate, and hopefully reveal, the manner in which that reversal takes place. One particular approach, that based on the operant conditioning phenomena (Blackman, 1974) appears to be particularly relevant to Action Selection. Other models of extinction (Balkenius and Morén, 1988) have investigated these phenomena in isolation. This paper considers and models the role behavioral extinction mechanism might play in the full context of multi-step Action Selection sequences.

Action Selection and behavioral extinction will be discussed in the context of Witkowski's *Dynamic Expectancy Model* (DEM) (Witkowski, 1997, 1998, 1999a, 1999b). The Dynamic Expectancy Model adopts a connectivist (Drescher, 1991; Witkowski, 1999b) approach. Essentially a method of building a network of rank ordered connections between current sense input and current goal, based on a continuously changing learned structure.

The Dynamic Expectancy Model selects actions on the basis of a function of the Animat's current sensory state and current goals. It reformulates that function (called the *Dynamic Policy Map*) dynamically as the motivations or "goals" of the Animat change with time, and as a consequence of what it learns as a result of the actions it selects.

In turn, the structure that underlies the generation of that dynamic function is built, and subsequently updated, with structural and tactical learning methods. As with other Action Selection models, the DEM is essentially an engineering artefact, but whose design principles are driven by our understanding of natural counterparts (Witkowski, 1997, for a detailed discussion). The approach here is not so much to provide detailed models of individual processes, but to investigate how these processes interact and the role they might play in animals and in artificial systems whose design is based on our understanding of animal behavior.

The biological inspiration that can drive our choices in designing artificial Animat systems (artificial Agents, based on biological principles) allows for the four Action Selection factors to be combined in a wide range of ways. Neural Networks emulate aspects of our understanding of

the brain's pathways to learn appropriate behaviors (Tani and Nolfi, 1998, for example). Reinforcement learning systems (Humphrys, 1998; Sutton, 1990 or Watkins, 1989) propagate the effects of occasional reward backward to create a *policy map* of sense-act pairs, ordered by current estimates of future reward. Classifier systems (Booker, *et al*., 1990; Riolo, 1991; Stolzmann, 1998) adopt a "bucket-brigade" approach to propagating the effects of occasional reward, and combine behavior selection with a genetic algorithm to create new classifier elements. Action Selection models may also be pre-defined, with little or no learning content (Maes, 1991; Tyrrell, 1993).

2 Behavioral Extinction

Behavioral Extinction (Blackman, 1974; Hergenhahn and Olson, 1993; Reynolds, 1968) describes the process by which a previously established learned connection is discarded when learned responses derived from it no longer elicit the desired outcome. Classical conditioning studies illustrate the reversibility of learning. Repeated association between *Unconditioned Stimulus* (US) and *Conditioned Stimulus* (CS) leads to the appearance and then gradual strengthening of the *Conditioned Response* (CR). Once established we note that the CR will weaken and apparently disappear following a period when US and CS are not associated, usually over a small number of trials (~10). Classical conditioning has been extensively modelled (Balkenius and Morén, 1988), who describe these extinction results as "not very surprising". However, not all learning is equal, and adopting a different experimental regime produces very different extinction patterns.

The raw behavioral data for the extension to the DEM described in this paper is derived from work using experimental techniques developed by B.F. Skinner to investigate operant conditioning learning. In an apparatus, now almost universally referred to as the *Skinner Box*, certain learning phenomena in animals may be investigated under highly controlled and repeatable conditions. In a typical Skinner box apparatus the subject animal may operate a lever to obtain a "reward", usually a small food pellet. The equipment may be soundproofed to exclude extraneous signals and different arrangements can be adopted to suit different species of subject animal.

Typically, the animal will be taught to operate the lever to obtain the reward before the start of an experiment. Once the subject is conditioned in this manner various regimes can be established to record effects such as stimulus differentiation, the effects of adverse stimuli ("punishment schedules"), the effects of different schedules of reinforcement, and, of course, experimental Behavioral Extinction. Progress of the learned response may be automatically recorded in a trace (fig. 1) that shows the number (and/or strength) of the emitted response events over a period of time.

For instance, we might train a rat to press a lever to obtain a food reward. We would expect the rat to try the lever to obtain food when it is hungry. If the lever is subsequently disconnected from the food dispenser, how long will the rat continue to try? Under appropriate conditions it is apparent that the rat will continue to operate the lever to (unsuccessfully) obtain food for a very considerable period, albeit at a decreasing rate. We ask, what purpose does such persistence in behavior serve? We also note that the reduction in the rat's lever pressing activities is not uniform, but adopts a distinctive pattern. Why should this be?

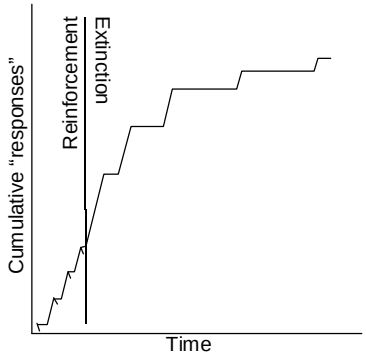

Figure 1: Operant Conditioning Behavior Extinction Curve

Figure 1 shows stylized experimental records (adapted from Reynolds, 1968) derived from Skinner Box experiments under an operant conditioning reinforcement schedule. The slope of the curve indicates the rate of the learned action "responses" (each such action causes an upwards increment in the trace), horizontal sections indicate periods with no responses. The downward tick marks indicate when a reward has been delivered. The traces to the left of the vertical separator line show the response curves under continued "normal" (rewarded) conditions, those to the right the effect following complete cessation of reward. This form of cumulative event graph is used later in figure 7.

One might expect that a regularly "rewarded" behavior should take longer to extinguish that one that has been rewarded only sometimes. A stronger, more rewarded, connection should surely take longer to eradicate than a partially rewarded one. This is not the case; partially rewarded behaviors consistently persist longer. This is referred to as the *partial reinforcement extinction dilemma*. It is also the case that although a behavior has been apparently fully extinguished (under both classical and operant regimes), it will reappear following a period of rest - the *spontaneous recovery* phenomena. Extinction behaviors under the operant schedule also exhibit the *latent extinction* phenomena, in which behavior actions that are contingent on a subsequent, but extinguished, activity in a chain of actions are themselves suppressed. It will become clear that latent extinction is a natural consequence of the DEM approach, and will hardly need further consideration. We will also suggest a straightforward interpretation of the partial reinforcement extinction dilemma in the context of

the DEM, and postulate a simple mechanism for spontaneous recovery.

There are two highly significant differences in the extinction curves obtained under classical and operant conditioning regimes. First, the characteristic stepped form of the curve in the *extinction phase* of the operant experiments following the cessation of rewarding events, which is quite distinct from the typically smooth extinction curve obtained under a classical conditioning regime. Second, that the number of events required to complete the extinction phase is much higher in the operant case than under classical conditioning.

3 The Dynamic Expectancy Model

The following sections provide a summary of the DEM mechanism. This description will emphasize the connectivity inherent in the model. Previous treatments of the Dynamic Expectancy Model (Witkowski, 1999b, for example) have concentrated on a formalism to more precisely define the structures involved and the processes by which they change.

The DEM forms (via a process of *structural learning*) and maintains (*tactical learning*) a network of connections between two interface components, *Signs* and *Actions*. This network of connections is (semi-)permanent, its structure changing relatively slowly in response to structural learning. The relative strength of connecting links in the network is controlled by tactical learning. These learning processes are described later.

Signs both define and detect *situations* that can be recognised by the Animat. Signs form the interface to the sensory apparatus available to a physical Animat. Signs may connect directly to sensors, but may equally be compound items, conjunctions of elemental sensory inputs. At each cycle of the algorithm every Sign is either detected or is absent and so evaluates to *active* or *inactive* for that cycle.

Actions, define the activities the Animat may perform. Where Signs defined the interface to the sensory apparatus, actions connect the DEM to the Animat's physical actuators. Actions being performed are deemed active.

Every action, a, has associated with it an *action cost*. The action cost indicates the relative effort that will be required by the Animat to complete that action. Action costs may be expressed in any units (such as elapsed time or energy expended) that may be consistently applied across all the actions used by the Animat. Action cost will be used in the "cost estimation" process for goal directed Action Selection.

The DEM also maintains a memory of recent activations and their associated timings for both Signs and actions. Information held on these *activation traces* is used by the structural learning component to construct new links in the network of connections.

The main learned structural connection in the DEM "network" is the *µ-hypothesis link*, figure 2.

$$s' \wedge a \rightarrow^{t \pm \tau} s''$$

Figure 2: The µ-Hypothesis Link

Each of these µ-hypothesis links should be read as an *expectation* of the form "performing the action a in the immediate context ($\wedge$) of s' predicts the occurrence of the condition s'' at time *t* in the future". The time *t* is bracketed by $\pm\tau$, forming a range a times to generalise the prediction in the temporal domain ($\tau << t$) and to overcome the potential effects of sensor sampling aliasing. Any particular combination of context Sign (s') and action (a) can potentially predict many consequences, similarly a consequence condition (s'') can be predicted by many different pairings, at different values of *t*.

Any µ-hypothesis link is deemed active whenever both its context Sign (s') and action (a) are active simultaneously. A new prediction (p) is created and added to the *Prediction Trace* for every instance of an activated µ-hypothesis link. Note in particular that this mechanism is invoked for <u>all</u> µ-hypotheses links meeting these activation criteria at any given time. The presence of active predictions drives the corroboration process.

The model's current estimate of the overall predictive ability for each µ-hypothesis link, the strength of the predictive connection "→", is recorded in the *corroboration measure*, C_m, ($0 \leq C_m \leq 1$). C_m is updated by the tactical learning mechanism based on a comparison of predicted outcome saved on p against actual outcome at every opportunity that arises to do so. The corroboration measure is central to the construction of the Dynamic Policy Map.

3.1 Corroboration: Tactical Learning

For each active prediction, the corroboration measure (C_m) of the µ-hypothesis link responsible for the prediction is modified according to:

$$C_m = C_m + \alpha(1 - C_m) \qquad \text{(eqn. 1)}$$

where the prediction was successful, and

$$C_m = C_m - \beta(C_m) \qquad \text{(eqn. 2)}$$

where the prediction was unsuccessful. Predictions are discarded from the prediction trace once this step is complete.

The positive *reinforcement rate*, α ($0 \leq \alpha \leq 1$), defines the rate at which successful predictions will strengthen C_m. Similarly, the *extinction rate*, β ($0 \leq \beta \leq 1$), defines the rate at which C_m will be weakened by failed predictions. Where no prediction was made the value of C_m remains unchanged. Sequences of successful (or unsuccessful) predictions give rise to the familiar negatively accelerating learning curve,

the values being normalized such that C_m rises asymptotically toward 1.0 (or falls toward 0.0).

3.2 Corroboration as Reward

The DEM μ-hypothesis link has many similarities to the notion of the three-term contingency, used by Catania (1988) to express the fully discriminated (Skinnerian) operant class of stimulus, action and outcome contingent on reward. It is perhaps little surprise, then, that findings from operant conditioning experiments seem particularly relevant to the understanding of these mechanisms. However, the DEM builds on the conjecture that the proper interpretation of this triple is that of a prediction, and that the strength of the connection ("→") should depend only on the predictive performance of the unit.

A conventional view would hold that the strength of the connection should be related to some goal or task specific "desirability" of s''. By adopting the predictive view, strength changes can be made internally, just by seeing whether the predicted event did or did not occur, independently of reward or reliance on an external agent to indicate correctness. This leaves the μ-hypothesis link uncommitted to any particular goal; learning is not task dependent. Changes to C_m can be applied immediately the prediction is verified, and is therefore always attributed to the specific μ-hypothesis link responsible for the prediction.

The Dynamic Expectancy Model is one of a number of contemporary learning Action Selection models that are based on the explicit use of prediction (hence *expectancy* model) to drive the learning processes (Tani and Nolfi, 1998; Stolzmann, 1998).

4 DEM Action Selection Methods

At each execution cycle the Animat must have some action to perform. Normally, the Dynamic Expectancy Model operates in two distinct modes for Action Selection:

(1) Goal directed Action Selection, and
(2) Exploratory Action Selection.

Any Sign can be assigned a *priority*. These prioritized Sign nodes then become goals for the Animat. Many Signs can be given a goal priority in this way, but the Animat treats only the one with the highest priority as *top-goal*, and will select actions to achieve this goal. The goal setting mechanism may be programmed into the model, or, as in the case for the experiments described later, Sign priorities may be manipulated directly.

Whenever the Animat is in goal directed Action Selection mode, the DEM algorithm attempts to construct and maintain a *Dynamic Policy Map* (DPM) from which it may then select actions directly. In this mode of operation actions are selected on the basis of the current sensory conditions, by direct reference ("reactively") to the Dynamic Policy Map. The construction and use of the DPM is described later.

When such a goal Sign becomes active (that is, the Animat detects it), it is deemed satisfied, and its priority is automatically reduced to zero. When the top-goal is satisfied, the next highest priority Sign automatically becomes top-goal, a new DPM is constructed and actions are then selected to achieve this Sign.

Whenever no top-goal is set the system defaults to selecting actions in exploratory mode. A wide variety of possible exploration strategies have been proposed (Thrun, 1992), and the choice of strategy will be implementation specific for any particular Animat model.

Irrespective of how an action came to be selected, the learning mechanism continually learns from the activities of the Animat. It updates the state of knowledge (and hence how the Animat will react in future) after each action according to the tactical and strategic learning strategies previously described. This is a natural, and powerful, consequence of the internal prediction based corroboration method.

5 The Dynamic Policy Map

Whenever a top-goal is set, the DEM attempts to create a Dynamic Policy Map (DPM) to form an ordering over sequences of links from every other Sign in the net to the Sign currently acting as top-goal. The DPM is conveniently thought of as an interconnected "graph" temporarily superimposed over the network of μ-hypothesis links. Signs associated with individual μ-hypotheses links represent the nodes and actions embedded within individual μ-hypotheses the arcs.

5.1 Constructing the DPM

The DPM is created by a process of *spreading activation* (Maes, 1991) propagating throughout the network of μ-hypothesis links from the top-goal Sign, which acts as a "seed" point.

Each μ-hypothesis link has associated with it a *cost estimate*, C_e, value. This cost estimate is computed from the given action cost of the action, a, embedded in the link, the current C_m value for the link and a *fatigue measure*, f_m, associated with the action:

$$C_e \leftarrow (\text{action\_cost}(a) * f_m) / C_m \qquad \text{(eqn. 3)}$$

Consider a situation where the corroboration measure (C_m, eqns. 1 and 2) is simply p(number of successful predictions|total predictions) made by a μ-hypothesis link - the probability that the μ-hypothesis link predicts correctly. With $f_m = 1$, the cost estimate value C_e would then be reasonably interpreted as the total estimated cost for the average number of attempts that must be made with the given μ-hypothesis to achieve the transition between s' and s'' that it predicts. A similar interpretation may be placed on the case for C_m shown in eqn. 1, with the proviso that the

"averages" are now biased towards recent experiences, with less recent experiences discounted away.

The *fatigue measure*, f_m, is normally unity, but is incremented by some small amount (the *fatigue increment rate*, FIR) each time the action is made. However, during periods when the action is not used it slowly reverts to unity (at the *fatigue recovery rate*, FRR). Its effect is to artificially raise the cost estimate of a DPM link where the action is used frequently, and therefore make any path in the DPM that uses this link more "expensive" and so less attractive.

Each Sign in the network will acquire a *valence level*, *v*, indicating the number (*n*) of µ-hypothesis links that must be traversed to reach the top-goal Sign "node" by the path of least total cost. The current top-goal Sign has a valence level of zero, the $\mathcal{S}'$ Sign of any µ-hypothesis link that leads directly to the goal (i.e. where its $\mathcal{S}''$ = top-goal) a valence level of 1, and so on. The *policy value*, P_v, of any node $\mathcal{S}$ at level *n* in the DPM is then directly expressed as the sum of individual estimated costs ($(C_e)^v$) by:

$$P_v(\mathcal{S}) \leftarrow \min \left(\sum_{v=1}^{v=n} (C_e)^v \right) \qquad \text{(eqn. 4)}$$

The policy value for each Sign $\mathcal{S}$ implicated in the DPM is computed by adding the cost estimate for its transition to the cost of the path to its $\mathcal{S}''$ node. If a lower cost path is encountered, the spreading activation process is re-activated for that node to minimize path costs at higher valence levels. The method used to compute these least (estimated) cost paths to create the DPM is a simple variant of the standard breadth-first graph traversal algorithm (for example, Nilsson, 1980).

Construction of the Dynamic Policy Map is complete when there are no further µ-hypothesis links that can be implicated, and no further path cost minimization can occur.

5.2 Selecting an action from the DPM

Following construction of the DPM the Animat has an estimate of the total "cost" of satisfying the top-goal starting from any Sign $\mathcal{S}$ included in the policy map. If <u>any</u> currently active Signs are included as a node in the DPM, then the action $\mathcal{A}$ included in the µ-hypothesis link associated with the <u>active</u> Sign node with the <u>lowest</u> P_v is selected. This is the action at the start of a sequence requiring the lowest overall estimated effort to achieve the top-goal.

Where there is no intersection between the set of active Signs and nodes on the DPM, an exploratory action is selected. These exploratory actions will either:

(1) achieve the goal directly (by chance),
(2) lead to a situation where a Action Selection from the DPM may continue, or
(3) cause new µ-hypothesis links to be created, which in turn expands scope of the DPM.

The DPM is recomputed frequently, whenever the top-goal changes, new µ-hypotheses are formed, or existing ones have undergone sufficient additional corroboration to indicate that a different solution path may be preferable.

6 Strategic Learning

Prediction, or rather the failure to predict the occurrence of a Sign, drives the structural learning component of the DEM, which is responsible for forming new µ-hypothesis links. The opportunity to create new µ-hypothesis links is indicated by appearance for the first time of a ("novel") Sign or by the appearance of a known but unpredicted ("unexpected") Sign. Previously unencountered Signs trigger the *creation by novelty* method. The appearance of an unpredicted, but previously known, Sign invokes the *creation by unexpected event* method. Unexpected Signs are detected by comparing the active predictions to the active Signs and applying the method to any unpredicted residue.

In either method a new µ-hypothesis link may be constructed from the novel or unpredicted Sign as '$\mathcal{S}''$', and a Sign ($\mathcal{S}'$) and action ($\mathcal{A}$) drawn respectively from the recorded activation trace of values of past Signs and actions. In this way the model creates a new "hypothesis" "that $\mathcal{S}''$ is predicted by performing the action $\mathcal{A}$ in the immediate context of $\mathcal{S}'$, at time *t* in the future". The timing relationship (*t* and hence τ in fig. 2) is derived from their relative positions in the respective memory traces. Once formed any new µ-hypothesis link will be tested for validity by the corroboration method, and incorporated into any DPM that is subsequently constructed.

To limit the rate at which new µ-hypothesis links are created the user may specify a *learning probability rate*, λ, which determines the probability with which a new µ-hypothesis link will be formed given one of these opportunities to do so. The Dynamic Expectancy Model also defines methods for differentiating partially effective µ-hypotheses by making their component Signs more or less specific (and so creating new Signs), and also for removing ineffective µ-hypothesis links.

7 Modelling Extinction

The extinction mechanism in the DEM interprets the steep and flat components of the operant conditioning extinction schedule (figure 1) as alternating periods of explicitly goal directed behavior interspersed with periods of exploratory activity when goal directed-ness is temporarily suspended. The DEM extinction mechanism is controlled by four factors, which determine the rates and relative effort expended on these two activities:

(1) The *valence break point* (VBP),
(2) the *valence break point factor* (VBPF),
(3) the *goal recovery rate* (GRR) and
(4) the *goal cancellation level*, Ω.

VBP and VBPF control the duration of the periods of goal directed activity, the GRR the duration of the intervening exploratory periods. The goal cancellation level

specifies the maximum value the path policy value may rise to before the top-goal will be cancelled by extinction. The effect of the interplay between these components can be clearly seen in figures 6 and 8.

Determining the VBP: When the Dynamic Policy Map for a new top-goal goal is first constructed, the lowest available policy value (P_v) associated with an active sign is taken as a measure of the likely cost of reaching the goal. The VBP is determined by multiplying the initial P_v by the VBPF (VBPF > 1, typically 10). The multiplier value is selected to give the Animat ample opportunity to achieve the goal by direct use of the DPM, allowing a generous margin for failed Actions.

Using the VBP: If the top-goal has not been satisfied, and the current best P_v reaches the VBP value, goal directed behavior is temporarily suppressed. The VBP is again multiplied by the valence break point factor in preparation for any subsequent periods of goal-directed activity. Each time a blocked μ-hypothesis link fails the estimated cost of the step increases (at an exponential rate), and the number of failed actions required to reach the next VBP level will be decreased as a direct consequence.

Using the GGR: On reaching each break point, behavior reverts to exploratory actions for a period determined by the GRR. Actions taken during this period are referred to as *unvalenced* actions, to distinguish them from "ordinary" exploratory actions. On the first suppression the goal recovery rate is high, and behavior reverts to goal directed quickly after only a few unvalenced actions. On reaching each subsequent valence break point the GRR is reduced (in the current implementation by a factor of two) and so the number of exploratory actions during the unvalenced period increases.

Goal Cancellation: Eventually, the current P_v will exceed the predetermined goal cancellation level, Ω, and the priority of this unachievable top-goal is automatically reduced to zero, releasing the Animat from the continued obligation to pursue that goal.

If at any point during this procedure the top-goal is satisfied the extinction process is cancelled.

8 The Investigations

To illustrate the effect of normal Action Selection in the DEM, and the circumstances, and effects of the Behavioral Extinction process, we now describe a sequence of three investigations.

Investigation One will reprise on the normal learning and Action Selection behavior of a DEM controlled Animat faced with simple alternative Action Selection sequences.

Investigation Two emulates the full extinction procedure. In part 2 of this investigation we repeat the experiment, but this time allow the Animat the opportunity to discover a new effective action sequence path, demonstrating a primary role for the extinction process.

Investigation Three replicates the Extinction procedure, but with two possible points in the Action Selection sequence, both of which are blocked.

Figure 3 shows the simulated Animat environment. The test environment used is the due to Sutton (1990). In all these experiments $\alpha = 0.5$, $\beta = 0.2$, $\lambda = 1.0$, FIR = 0, $\Omega = 10^6$ and all actions cost are 1.0. Although restricted, this test environment allows considerable control over the experimental conditions and allows for straightforward analysis of the results (and in particular, easy and clear visualization of the DPM). Following Sutton, a random exploratory strategy is adopted. While less efficient that alternatives, its use eliminates a potential source of domain specific bias in the investigations.

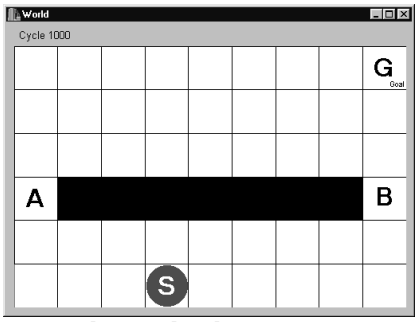

Figure 3: The Simulated Animat Environment

8.1 Investigation One

In this investigation the Animat is placed at the start location "S" and allowed to explore the environment by a random walk strategy for 1000 steps. This is sufficient to ensure that the environment is fully explored, and that both possible paths to "G" have been learned. No goal is asserted during this initial period. The Animat is allowed to learn *ad libitum*, and so μ-hypothesis links will be created at every opportunity and the environment learnt (adequately for the investigation, although not completely at this stage). Learning is not contingent on an external source of reward, as corroboration is an entirely internal process. This is *latent learning* (Witkowski, 1998). Learning has occurred, but will not be made manifest until a motivating goal is set. The ability to perform latent learning is often seen as a primary differentiator between this class of learning mechanism and true (reward based) reinforcement learning methods.

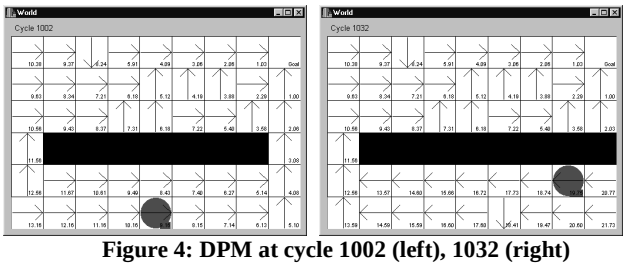

Figure 4: DPM at cycle 1002 (left), 1032 (right)

At the conclusion of the 1000 cycles of unrewarded and unmotivated exploration, the Animat is returned to the start

location "S" and the location "G" assigned a goal priority value of 1.0, making it top-goal. As the top-goal has been modified, the system constructs a Dynamic Policy Map using "G" as the seed point (fig. 4, left). Animat actions are now selected from the DPM in a reactive manner until the goal location is reached, much as they would be from a static policy map developed by a *Q*-learning method (Sutton, 1990, Watkins, 1989). Unsurprisingly the Animat selects the shorter of the available paths, and moves towards the goal location. Once the goal location is reached (after 10 steps) the goal is automatically cancelled and behavior would revert to exploration.

In the next stage of the investigation the Animat is returned to "S" and "G" is again made top-goal. An obstacle is placed at location "B", blocking the shorter path. The Animat attempts to follow its best path, but is prevented by the obstacle (note that the Animat is unable to sense the obstacle until it is encountered). Each failed attempt to apply the µ-hypothesis link (identified as "H14" in fig. 5), responsible for the blocked transition, increases the estimated cost of that transition according to the extinction rate β. Each time the DPM is recomputed this additional cost increases the path cost by that of the failed µ-hypothesis link. At some point the cost of the best alternative path in the DPM is exceeded (12 attempts in this example), and the Animat switches path to traverse the environment via location "A" (fig. 4, right).

Figure 5 shows the total estimated remaining path cost (triangles) and the cost contribution to that total by µ-hypothesis "H14" (squares). The first segment, from step 1000, shows the cost falling steadily as the shorter path is traversed. The second segment, from step 1012 shows the estimated cost falling until the block at "B" is encountered, then rising until it exceeds that of the longer path via "A", and falling again until the goal is reached (DPM of fig. 4, right).

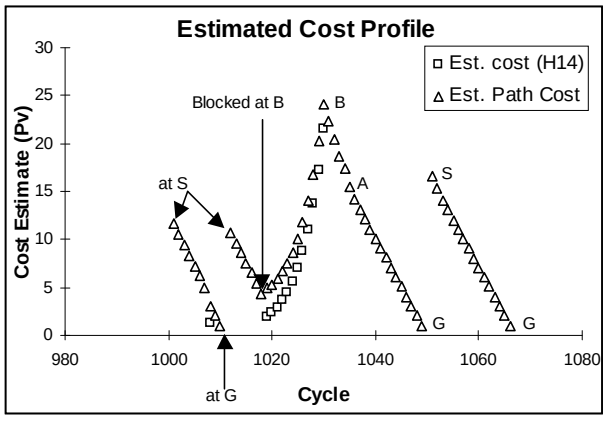

Figure 5: Estimated Cost Profiles

In the last part of investigation one the Animat is returned to "S", "G" set again and the Animat released. It will be apparent from fig. 5 (cycles 1051 – 1068) that the Animat immediately traverses the path via location "A"

8.2 Investigation Two

This investigation determines the goal extinction behavior of the Animat when a single, previously established, path is obstructed. These conditions replicate the Skinner box experiments, but in the context of a multi-step action-sequence. Location "A" is blocked, then the Animat is allowed to explore the environment for 1000 steps, as investigation one. The Animat is returned to "S" and the goal location "G" established as top-goal. We confirm that the expected direct path via "B" is taken. Now the Animat is again returned to "S" and "G" set again as goal. Before the Animat is released an obstacle is introduced at location "B", such that there is now no possible route to the goal location.

Figure 6 monitors two important internal parameters during the extinction process. The estimated cost of the goal path (square markers) and the valence break point (VBP) value (circle markers). The DPM is computed, and a path cost estimate of about 10 units derived (ten steps, each of action cost one). With a valence break point factor of 10, the valence break point is therefore set to 100. As there is no alternative method to reach the goal the Animat persists with the only µ-hypothesis link available, and the cost estimate rises as previously described due to successive failed corroboration steps.

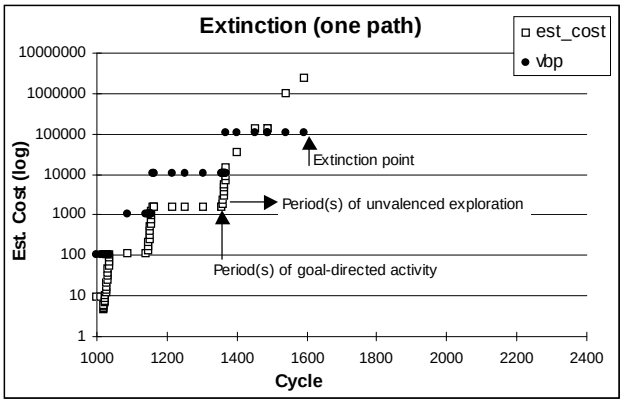

Figure 6: Single path to extinction cost estimates

If no alternative path is encountered before the estimated path cost exceeds the current VBP (as it would have in investigation one), the Animat reverts to exploratory actions for a period determined by the goal recovery rate. Initially the recovery rate is high. Following 101 steps of unvalenced activity, goal-directed behavior is restored and a new DPM computed. The estimated path cost is again increased by the valence break point factor (to 1000) to give a new VBP. The Animat immediately returns to the only known path and attempts that until the cost again exceeds the new VBP value. Exploratory actions are resumed, with a reduced goal recovery rate, so exploration continues for longer.

This procedure is repeated with increasing periods of exploratory activity punctuated by decreasing periods of goal-directed activity. At some point a (failed) corroboration of the blocked µ-hypothesis will result in the total path cost

estimate exceeding the goal cancellation level, Ω. This is the *goal extinction point* and the goal is cancelled. The goal might be reasserted, but to little useful purpose.

Some Skinner Box extinction curves record the bar being pressed during the periods of apparent inactivity (equivalent to unvalenced exploration). The Model shows a similar effect, the conditions to activate the blocked µ-hypothesis link may occur at any time, and so the extinction point can be reached at any point in the experiment. Although they record quite different activities, the estimated cost record of figure 6 can be viewed as an analogue for that in figure 1. Each marker square represents one action at the critical blocked point

8.3 Part 2: The Blocking Procedure

The procedure here replicates Sutton's (1990) *blocking procedure*, and differs in some details to that used previously in investigation two. Sutton uses the blocking procedure to investigate his *exploration bonus* method for mixing exploration and exploitation of the environment in the Dyna-*Q*+ algorithm. Under this procedure initial exploration assumes that "G" acts as a source of reward (goal) continually. Path "A" is blocked initially (one path to "G" via "B"). The Animat is placed at "S", and "G" is made top-goal (it was not valenced in investigation two, but without latent learning, if this were not the case now the *Q*-learning mechanism of Dyna-*Q*+ would learn nothing during this phase). Each time the Animat reaches "G", one unit of reward is noted, and the Animat returned to "S". "G" is always immediately reinstated as top-goal. This procedure is repeated up to step 1000.

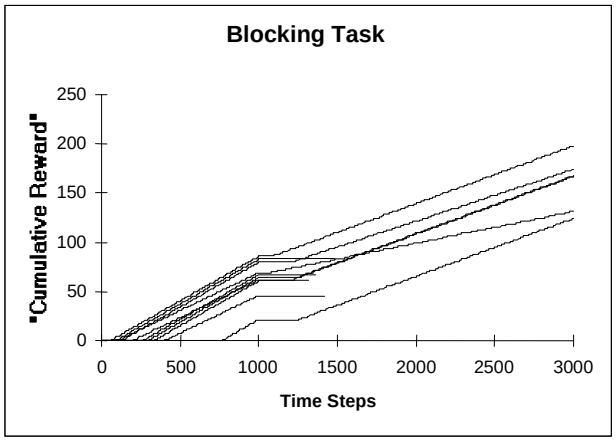

Figure 7: Blocking Task, Individual curves

At this point an obstacle is placed at "B" (as in part 1), but the obstacle at "A" is removed. The procedure now continues until cycle 3000. An Animat must discover the new path (during a period of unvalenced activity), before continuing to receive reward at "G" (and being automatically returned to "S" each time).

Figure 7 shows the performance of ten individual DEM Animats under these conditions. These "individuals" are in effect "clones" differing only in the starting seed used to generate the random Action Selection during the exploration phases. The slope of each line indicates the frequency with which the Animat reaches the goal and receives one unit of reward. The steeper the slope the more quickly the Animat has reached the goal location. Flat portions from step 0 represent times when the Animat is exploring the environment as the initial random walk. The variability and delay shown is typical for the random walk employed. From step 1000 the Animats enter the extinction process described previously.

Figure 7 shows that six of the test DEM individuals successfully located and traversed the new path at location "A", while four failed to do so and suppressed the goal. These individuals were withdrawn from the experiment. If they remained in the environment these Animats would continue with exploratory actions. Clearly the speed at which the Animat might connect with the (previously explored) upper part of the environment, and the chance of extinction is determined to a great extent by the nature of the exploration strategy adopted. Sutton (1990) reports that after a period of re-exploration of about 500 steps Dyna-*Q*+ locates the new, but previously unknown, path and continues to garner reward. The results are broadly comparable, but the DEM adopts a biologically inspired strategy, where Sutton makes his choice on the basis of his method's numerical properties.

8.4 Investigation Three

This final investigation repeats the extinction experiment of investigation one, but in a situation where the Animat has two paths available ("A" and "B") from start to goal during the initial 1000 step (no goal set) exploration phase. Then both paths are blocked before starting the extinction phase at step 1000. The Animat behavior is modified to appearing to scuttle back and forth between the two previously effective paths during periods of goal-directed activity interspersed with random exploration during the unvalenced periods.

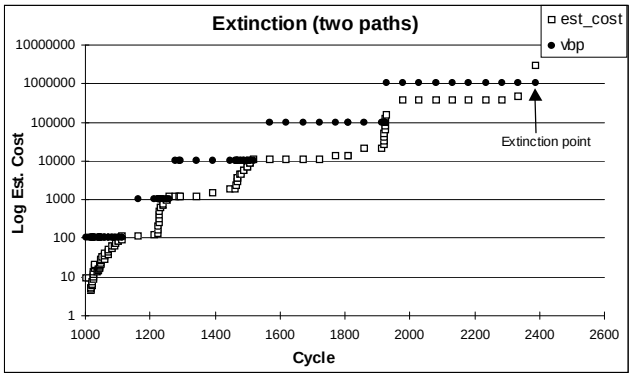

Figure 8: Dual Path Extinction

Figure 8 shows the resulting estimated cost and VBP values for this investigation. Figure 9 shows the detailed effect of this scuttling behavior during the first period of goal-directed activity. Each increase in the cost estimate arises from the Animat attempting a blocked µ-hypothesis link, first at one end of the maze, then at the other. The Animat appears decreasing persistent in its attempts to traverse each of the known blocked paths with each successive attempt. Gaps between the cost estimate rises indicate those cycles during which the Animat is (under goal directed control) travelling between the two places where the known paths had been located.

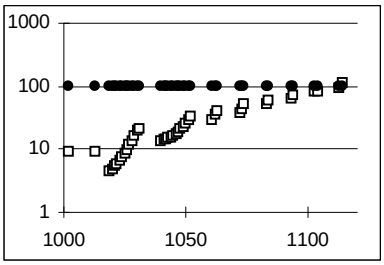

Figure 9: Dual Path Extinction (Detail)

9 Summary

This paper has described a behavior extinction mechanism based on experimental data for the Dynamic Expectancy Model. The effect of this mechanism is to provide the last stage in a series of behavioral shifts by which the Animat may manage changes to and failures of Action Selection. The various investigations reporting experiments with the Dynamic Expectancy Model's implementation, SRS/E, confirm the sequence of events:

1) If the Action Selection function defined by the Dynamic Policy Map is effective, then the Animat achieves its top-goal, and all is well.
2) If the Action Selection function fails to achieve the top-goal and there is an alternative, but initially less desirable route, the Animat will attempt the first route until the estimated policy value exceeds that for the alternative, which is then pursued. If the second was blocked a third might be attempted, and so on. One might assume that this is the normal state of affairs, a combination of trying alternatives and persistence eventually leading to a successful outcome.
3) If no viable alternative exists, the Animat will continue to attempt the failed path, or paths, until the estimated cost exceeds the previously calculated valence break point (VBP). This initiates a brief period of (unvalenced) exploration, before the Animat returns to goal-directed Action Selection behavior on the only known path, but with a higher value for VBP.
4) If a fresh path is discovered during one of these periods of exploration, new µ-hypothesis links will be created and a new Action Selection path to the top-goal may become available. This was noted in investigation 2, part 2.
5) This alternating explore/goal-directed behavior continues until the goal extinction point is reached, when priority of the offending top-goal is forcibly reduced. The Animat may then revert to other, hopefully more productive, behaviors, (investigations 2 and 3).

The extinction mechanism provides a behavioral escape mechanism "of last resort", and prevents the Animat from being forced (by its own Action Selection mechanism) to display increasing desperate and inappropriate behavior, possibly to its severe detriment. It is not a panacea, as although the "hungry" Animat is no longer forced to follow the goal-directed path to known sources of food it is, unfortunately, no less hungry.

10 Discussion and Conclusions

Throughout this paper we have attempted to show how findings from experimental procedures elucidating the form of the behavioral extinction can be integrated into a current model of Animat learning and behavior. We have investigated the role it plays in the overall strategy of an Animat.

The form of the extinction mechanism is clearly related to the explore/exploit trade-off problem (Thrun, 1992; Wilson, 1996). Reinforcement learning systems must balance the time they spend moving towards known sources of reward with actions that may encounter new, and better, sources of reward. In general, the DEM does not need to interleave exploration actions when it is pursuing a goal-directed behavior, it explicitly takes its "best" possible route to the top-goal. Normally, DEM Action Selection behavior reverts to exploration at times when there is no top-goal. Indeed, it would be expected to encounter novel situations (and learn from them) while pursuing a variety of different goals as it goes about its everyday activities. During Extinction, these lengthening periods of unvalenced exploration progressively widen the search space for new solution paths. Periodically, but with less vigor, the Animat periodically returns to its known "best" path, as this remains its best option to achieve the current top-goal.

Investigation 3 presents results from a situation where two known paths are extinguished. The behavior of the Animat in these circumstances is quite distinctive. It would be instructive to repeat the procedure under laboratory conditions with animal subjects, as this information is not directly recorded in the Skinner box experimental paradigm.

The findings of Investigation 3 provide an effective interpretation of the *partial reinforcement extinction dilemma*. The observation that quickly formed or fully rewarded behaviors can be extinguished more rapidly than those that develop under conditions of partial reinforcement. We note the difference between extinction time in investigation 2 (single path, average of 10 runs = 870.9steps) compared to those in investigation 3 (dual path,

average = 1443.2 steps). Two factors are involved here. First, the extinction mechanism does not take effect until the contribution of the worst of the alternative μ-hypothesis links has reached the appropriate break point. Second, to reach that point, each of the alternative actions must be tried, typically many times. In the case of investigation 3, this was exacerbated by the extra actions required to travel between the two alternatives.

With a single point of "reward", if the Animat creates an effective μ-hypothesis link, and the prediction it makes (that the reward will occur) always succeeds, then the structural learning mechanism will not be invoked. Where these predictions fail, new μ-hypothesis links can be created, leading to the formation of alternatives to the reward. The more alternatives, the longer it takes to extinguish them. This explanation has yet to be tested under equivalent experimental conditions, but it is substantially more straightforward than those proposed that assume only a single reinforced connection.

Consider the effect of the fatigue measure (f_m) of eqn. 3. The Extinction mechanism ensures that one or a small number of actions are attempted at an elevated rate. This will cause the value of f_m to rise proportionately. In turn this increases the effective policy value for the path, causing the various valence break points to be reached prematurely. The extinction point is therefore also reached sooner than it would be on the basis of corroboration alone. Following extinction, the selection rate for the action(s) falls, and the fatigue recovery process starts. After a suitable period of recuperation the policy value falls back below the extinction point and the behavior appears to *recover spontaneously*.

The Spontaneous recovery phenomenon indicates that whatever strategy an Animat might have for removing ineffective μ-hypothesis links under normal circumstances, it appears remarkably tenacious in retaining at least the last link option in a previously successful chain. Presumably having a bad solution to a problem is better than having none at all. In any case circumstances may well revert to a situation where the link works again.

11 References

Balkenius, C. and Morén, J. (1988) Computational Models of Classical Conditioning: A Comparative Study, 5th Int. Conf. on Simulation of Adaptive Behavior, pp. 348-353

Blackman, D. (1974) Operant Conditioning, Methuen & Co.

Booker, L.B., Goldberg, D.E. and Holland, J.H (1990) Classifier Systems and Genetic Algorithms, in: Carbonell, J.G. (Ed.) Machine Learning: Paradigms and Methods, The MIT Press, pp. 235-282

Catania, A.C. (1988) The Operant Behaviorism of B.F. Skinner, in: Catania, A.C. and Harnad, S. (eds.) The Selection of Behavior, Cambridge University Press, pp. 3-8

Drescher, G.L. (1991) Made-up Minds: A Constructivist Approach to Artificial Intelligence, The MIT Press, Cambridge, MA

Hergenhahn, B.R. and Olson, M.H. (1993) An Introduction to Theories of Learning, Prentice Hall, New Jersey

Humphrys, M. (1998) Action Selection Methods using Reinforcement Learning, 5th Int. Conf. on Simulation of Adaptive Behavior, pp. 135-144

Maes, P. (1991) A Bottom-up Mechanism for Behavior Selection in an Artificial Creature, 1st Int. Conf. on Simulation of Adaptive Behavior, pp. 238-246.

Nilsson, N.J. (1980) Principles of Artificial Intelligence, New York: Springer-Verlag (Symbolic Computation Series)

Reynolds, G.S. (1968) A Primer of Operant Conditioning, Glenview, IL: Scott, Foresman & Co.

Riolo, R.L. (1991) Lookahead Planning and Latent Learning in a Classifier System, 1st Int. Conf. on Simulation of Adaptive Behavior, pp. 316-326

Stolzmann, W. (1998) Anticipatory Classifier Systems, 3rd Annual Conf. on Genetic Programming, pp. 658-664

Sutton, R.S. (1990) Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming, Proc. 7th Int. Conf. on Machine Learning, pp. 216-224

Tani, J. and Nolfi, S. (1998) Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems, 5th Int. Conf. on Simulation of Adaptive Behavior, pp. 270-279

Thrun, S.B. (1992) The Role of Exploration in Learning Control, in: White, D.A. and Sofge, D.A. (eds.) Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches, Florence, KY: Van Nostrand Reinhold 41022, 27pp.

Tyrrell, T. (1993) Computational Mechanisms for Action Selection, University of Edinburgh, Ph.D. thesis

Watkins, C.J.C.H. (1989) Learning from Delayed Rewards, King's College, Cambridge University, Ph.D. thesis

Wilson, S.W. (1996) Explore/Exploit Strategies in Autonomy, 4th Int. Conf. on Simulation of Adaptive Behavior, pp. 325-332

Witkowski, M. (1997) Schemes for Learning and Behaviour: a New Expectancy Model, Dept. Comp. Sci., Queen Mary Westfield College, Univ. of London, Ph.D. thesis. (http://www.ee.ic.ac.uk/mark/slab.htm)

Witkowski, M. (1998) Dynamic Expectancy: An Approach to Behaviour Shaping Using a New Method of Reinforcement Learning, 6th Int. Symp. on Intelligent Robotic Systems, pp. 73-81

Witkowski, M. (1999a) Applying Unsupervised Learning and Action Selection to Robot Teleoperation, Proc TIMR-99, Towards Intelligent Mobile Robots, 9pp.

Witkowski, M. (1999b) Integrating Unsupervised Learning, Motivation and Action Selection in an A-life Agent, 5th Euro. Conf. on Artificial Life, pp. 355-364