Chapter 6

6. Investigations and Experimental Results

This chapter describes a series of experiments with the SRS/E program. The approach has been to investigate the properties of the algorithm under highly controlled conditions, allowing a clear view of the algorithm's behaviour and performance. Some of the investigations mirror those used to investigate reinforcement learning systems from the modern machine learning paradigm, but some revive and repeat historical investigations used to disambiguate between competing theories of natural learning. It is interesting to note that these issues are still debated as actively as ever after decades of research. There are significant differences in the constitution of animals and animats, and some of the procedures must be modified to reflect these. Nevertheless it is hoped that the spirit of the original experiments is faithfully captured, and some of the lessons and challenges revealed will make a substantive contribution to this ongoing debate.

The previous chapter described the provisions that have made to enable the investigator to design and conduct experiments with the SRS/E program and to analyse and present the results obtained. Section 6.2 of this chapter describes a series of "baseline" experiments in which the performance of the SRS/E algorithm is compared directly to the performance of the Dyna-PI algorithm described by Sutton (1990). The SRS/E algorithm performs the task described by Sutton more efficiently by a factor of some 40 times. Additional investigations in this section clearly demonstrate the development of the classical negatively accelerating learning curve from the widely varying performance of many individual animats, in a manner predicted by the *stimulus sampling theories* previously mentioned.

Experiments described in section 6.3 determine the effects of "noise" on the performance of the SRS/E algorithm. These experiments adopt a definition of noise provided by Sutton, and clearly indicate that the SRS/E algorithm will learn

effective solutions even when presented with high levels of disruptive noise. These experiments also distinguish between the effects of noise on the learning process and on animat behaviour. Direct comparison with the Dyna-PI algorithm was not possible as Sutton did not report results with his algorithms.

The experiments described in section 6.4 investigate how the SRS/E algorithm responds to multiple and alternative goals. A number of experimental situations are explored which demonstrate the flexibility provided by the Dynamic Policy Map approach adopted by the SRS/E algorithm. In the alternative goal experiments the animat is required to traverse between a known start and goal situation, which is then reversed (such that the start becomes the goal and *vice versa*). In the multiple goal experiments the animat must visit several, arbitrarily selected goals. These tasks are not achievable with an unmodified *Q*-learning algorithm or any of Sutton's Dyna algorithms, as they all use a static policy map, and so no comparison of performance can be possible. These experiments therefore highlight a radical improvement between existing external reward and the Dynamic Expectancy based methods of reinforcement learning introduced by this thesis.

The investigations described in section 6.5 replicate experimental conditions used by Sutton to determine the effects of blocking known solution paths and opening new solution paths during individual trials of his Dyna-Q+ algorithm. Dyna-Q+ is a specifically modified variant of the Dyna-PI algorithm to address these tasks. The SRS/E algorithm matched the published performance in all the tasks described, although the method employed by the two algorithms is substantially different. SRS/E incorporates an extinction mechanism, not present in *Q*-learning or the Dyna algorithms, which allows the animat to abandon unachievable goal directed tasks and thus escape from potentially "life" threatening situations. The extinction mechanism is developed on biologically plausible grounds.

The experiments of section 6.6 replicate classic "latent learning" procedures. The latent learning experiments were the first to demonstrate conclusively that learning in animals could take place in the absence of external reward or reinforcement. Latent learning may be easily demonstrated with the SRS/E algorithm, and this chapter replicates the procedures adopted to show the effects in animal experiments. Demonstration of latent learning by a reinforcement algorithm

employing the *Q*-learning or Dyna methods would appear to be highly problematic, and remains a challenge to those espousing that school of thought. Similarly section 6.7 describes a replication of the "place learning" experiments, in which the animat must make different responses when placed in apparently identical stimulus situations from trial to trial. While the SRS/E algorithm responds to the place learning challenge in a similar manner to experimental animals, it remains unclear how a conventional reinforcement algorithm based on a static policy map could achieve this.

It might be noted that Sutton was obliged to employ a family of algorithms, Dyna-PI, Dyna-Q and Dyna-Q+, to demonstrate the experimental procedures described in this chapter. A single program implementing the SRS/E algorithm has been used for the experiments to be described.

6.1. The Individual Experiments

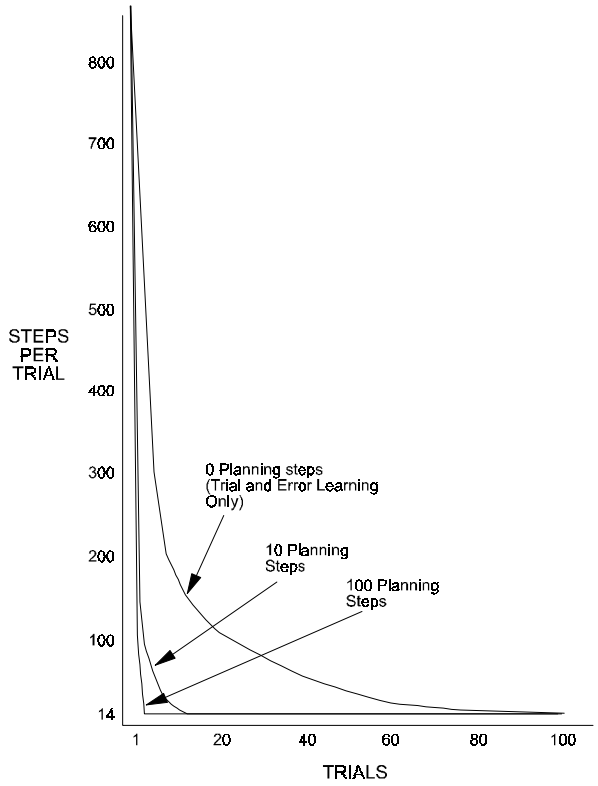
The sections that follow describe a series of individual experiments that attempt to characterise the performance of the SRS/E algorithm in well defined and controlled environments with particular reference to its learning capabilities. Each section is divided into three major parts. Part one will consider the rationale for the experimental schedule and describes the method and experimental procedures adopted for the experiment. As these may be derived from two separate methodologies, natural learning and machine learning, some care will be taken to ensure the data is extracted appropriately to identify and accommodate cross-domain issues. Part two will present the results from specific experiments. Wherever possible this presentation of results will take graphical or tabular form to provide for easy assimilation of the main points being investigated. Where a comparative investigation is being performed (one which replicates or substantially adapts part or all of an established procedure) an attempt will be made to present the SRS/E results in a form reflecting that of the original or source work, where this does not unduly impact or compromise the current experiments. Part three discusses the results of the experiment.

6.2. Baseline Investigations

These initial experiments attempt to characterise the SRS/E algorithm under highly controlled conditions, and to compare its performance to a well-established example of reinforcement learning. Sutton (1990) has extensively investigated a family of algorithms related to the idea of dynamic programming. To establish a performance baseline SRS/E is tested under conditions functionally identical to the descriptions given for Dyna-PI and "learning curves" (indicating improvement in performance following practice) obtained. Dyna-PI is presented by Sutton as showing substantial performance improvements over previous reinforcement learning methods.

Dyna-PI alternates "actual" movements in its simulated environment with "hypothetical experiences" derived from a world model created from data gathered during the actual exploration phases. Sutton refers to these periods of hypothetical activity as "planning"; a more apposite term might be "rehearsal". The three curves of figure 6-1 indicate the effect of increasing the ratio of "hypothetical experience" relative to "actual experience". The outer curve, labelled "0 planning steps" is equivalent to the performance of the underlying learning algorithm, converging with the optimal performance line (14 steps/trial) after about 90 trials. Where the animat is permitted 10 "planning" steps interspersed with each actual trial the curve reaches the optimal value after some 12 trials. As the ratio increases, the performance improvement becomes ever more apparent. In effect an equivalent amount of computation has been performed, although observable activity is substantially reduced.

SRS/E retains no additional internal world model. To obtain baseline learning curves SRS/E will be successively handicapped by artificially limiting the frequency with which it can exploit a recognised learning (by creation) opportunity. This is achieved by manipulating the learning probability rate (Lprob), while leaving other experimental conditions unchanged. Varying the learning probability rate introduces sampled learning, partially emulating the effects of spurious or irrelevant signs being incorporated into μ-hypotheses.



\monolith\mazes\graphic 5.4

Figure 6-1: Results from Sutton's Dyna-PI Experiments
(from Sutton, 1991, p. 219)

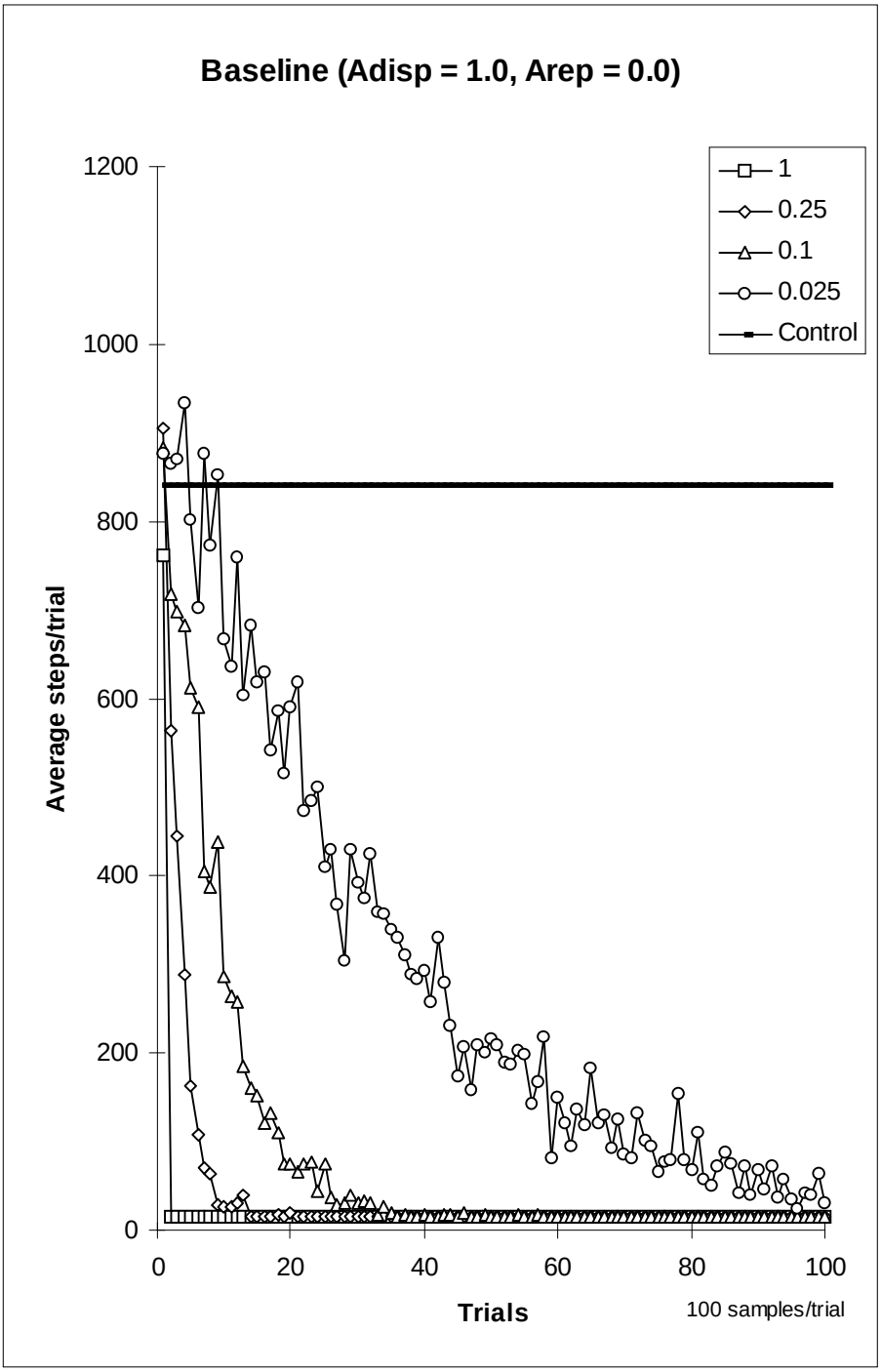
6.2.1. Description of Procedure

To perform the baseline experiments the first fixed schedule is used, which automatically selects and initialises the DynaWorld/Standard environment. Four separate learning curves are created with four different values of the *learning probability rate*, 1.0 (all learning opportunities taken), 0.25 (25% of opportunities taken), 0.1 (10% of opportunities) and 0.025 (2.5% of opportunities). The other factors are held constant for the duration of the experiment. In addition a control baseline is established indicating the animats' performance without valenced behaviour. Each curve is the average of 100 separate experimental runs, each of 100 trials. For each run a new animat (based on a new random starting seed) is placed at the starting point (located at X = 0, Y = 3) and allowed to run the maze. The number of steps taken to reach the goal (at X = 8, Y = 5) are recorded for each trial.

At the conclusion of each trial the animat is returned to the starting point, the goal reasserted (with a priority of 1.0) and the animat released to traverse the maze following whatever valenced path is available. In Sutton's experimental paradigm reward is assigned and the animat is returned to the starting location when the goal is reached. As corroborative learning does not take place in SRS/E until predictions are verified, the animat is allowed to remain undisturbed in the experimental maze for an additional 16 execution cycles after the goal is reached before the trial ends. Each curve is therefore composed of 10,000 visits to the goal location (100 runs of 100 trials). The control line is determined from 2,500 random walks from start to finish. The complete experiment comprises 42,500 visits to the goal location. This is comparable to Sutton's experimental design. The remaining system and animat parameters were held constant throughout the procedure (Arep = 0.0, Adisp = 1.0, $\alpha = 0.5$, $\beta = 0.2$, $\gamma^1 = 0.0$, $\gamma^2 = 0.9$, $\gamma^3 = 0.1$, $\gamma^4 = 0.0$).

6.2.2. Results and Analysis of Baseline Experiment

Figure 6-2 summarises the results of the baseline learning experiments. With learning probability rate = 1.0 every opportunity to learn by creation is taken. As the exploration by random walk is protracted due to the selection of a new random action at each cycle most of the possible μ-hypotheses have been created by the first time the goal location is reached. The random walk length for the first trial is highly variable (average of the 100 runs 743.25, best 24 steps, longest 4,380). On being returned to the starting point for a valenced trial to the goal location there is consequently a good chance that an optimal (there are many such paths), or nearly optimal path will be created. The average path length for this second trial is 15.32 (best is 14 steps). Of the 100 runs, 53% of the second trial achieved the optimal 14 step path, 34% the 16 step path, 8% the 18 step path, 4% the 20 step path and one path of 22 steps. By trial 100 the average valenced path length had fallen to 14.96, still above the achievable best.



monolith\results\bse1all\base1.xls

Figure 6-2: Baseline Learning Curves (Lprob = 1.0, 0.25, 0.1 and 0.025)

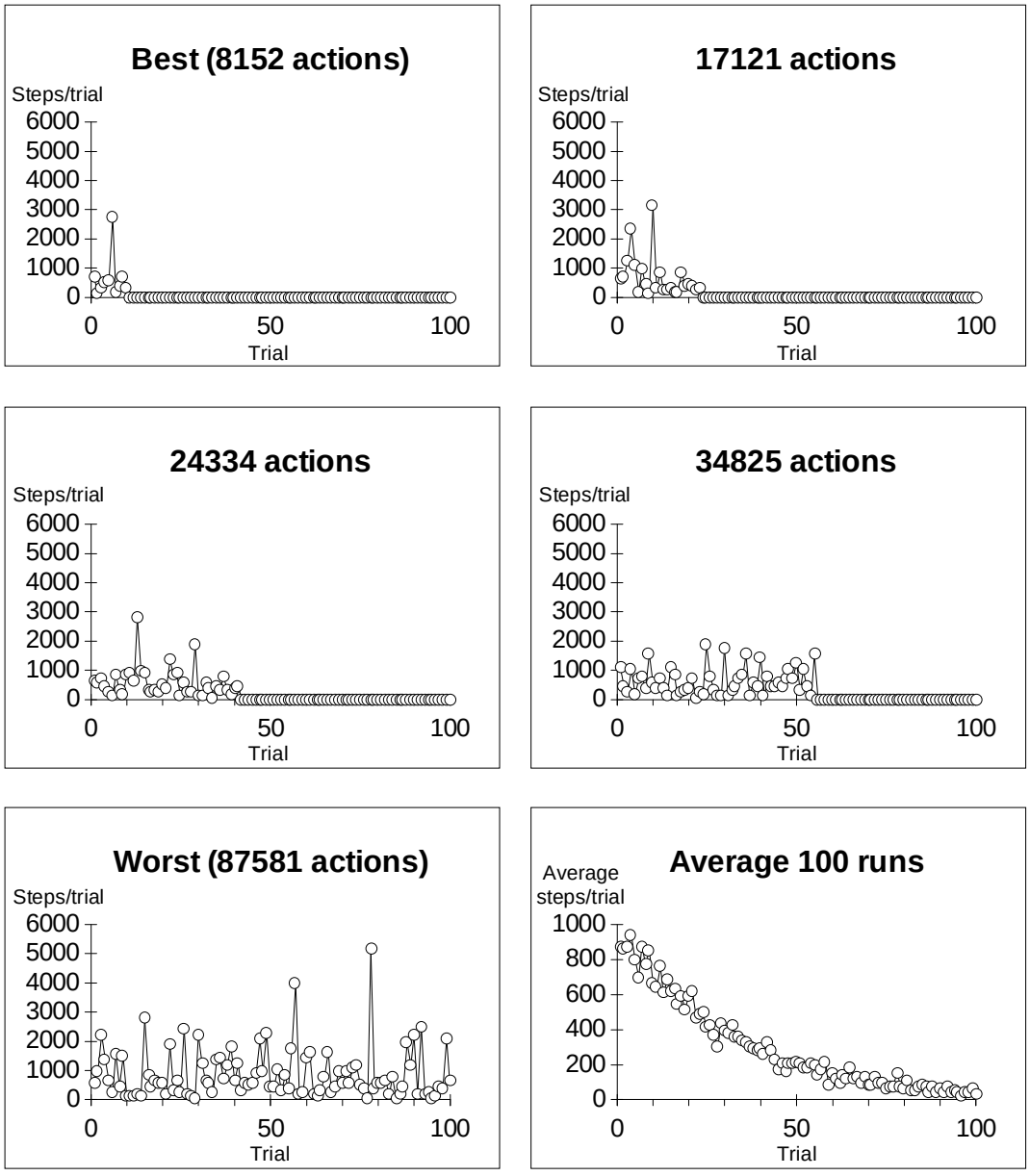
With values of Lprob less than unity, the learning curves take on a more traditional appearance. Discovery of the optimal (or near optimal) path is delayed. The effect of decreasing the probability that a learning by create event will occur has a quite distinctive effect on the rate at which performance improves (as indicated by falling

steps/trial), and on the point at which performance stabilises at its minimal level. The last animat to find its stable valenced path for Lprob = 0.25 (diamond graph markers) is at trial 26, the last one for Lprob = 0.1 at trial 56 (triangle markers). The penultimate animat for Lprob = 0.1 stabilised at trial 40. This point of stability has not been reached for the Lprob = 0.025 curve after 100 trials, four individuals from the initial 100 animats still not having found a complete valenced path. An individual animat is defined here as an animat assigned a specific value to the pseudo-random number generator seed (*rseed*) at *parturition*. This value will remain unchanged for the individual for the duration of the experiment.

Figure 6-3 details the performance of a selection of individual animats from the Lprob = 0.025 curve. The five individuals are selected on the basis of the total number of actions they took during the experimental run. Individuals were ordered according to the total number of steps taken in the 100 trials, the sub-figures indicate the "best" (fewest steps), the "worst" (most steps) and the quartile individuals. The "best", individual 84, (rseed = 840) made a total of 8,152 actions (minimum possible is 1,400, figures exclude the run-on period), stabilising by trial 11. Individual 69 (rseed = 690) had stabilised by trial 10, but the preceding random walks had taken more steps. The individual ranked 25th in the population (individual 68) stabilised on trial 24, 50th (individual 78) at step 42, 75th (individual 9) on trial 56 and the "worst" (individual 99) finally stabilised on trial 116. The net effect is shown in the lower right sub-figure.

For each trial, where Lprob ≠ 1.0, the transition from a poor solution path to the near optimal, stable, one is in most cases quite distinctive and often abrupt - as though "the penny dropped". Inspection of the trace information confirms that the effect is primarily due to the probability with which μ-hypotheses at low valence levels leading to the goal sign are formed. Until these particular μ-hypotheses have been created the formation of an effective Dynamic Policy Map is not possible, and so the majority of actions remain unvalenced. Even though this final step is not in place the learning of other μ-hypotheses is still taking place. Once the near goal connections are made, with a probability regulated by Lprob, sufficient μ-hypotheses are invariably available to create an effective DPM from start to goal. A less common effect where a short "stub" DPM builds out from the goal, which subsequently connects to the main body of knowledge is also observed. The overall

observable effect on measured path lengths of this stub phenomenon in relation to random walk length is small. This interpretation of the probabilistic nature of the learning process has much in common with the *stimulus sampling theories* promoted by *William Estes* and others (Bower and Hilgard, 1981, Ch. 8 for summary of this position).



monolith\figures.ppt:slide 3

Figure 6-3: Contribution of Individual Animats to Learning Curve

6.2.3. Discussion

Under the learning conditions defined by the learning curve where Lprob = 1 the performance comparison with Sutton's Dyna-PI system is clear. Where Dyna-PI takes approximately 90 trials to reach a stable minimum path solution, SRS/E does so in a single trial across all individuals in the test population. Dyna's poor performance in these circumstances arises from two properties. First, reinforcement is only made at the point the animat reaches the goal, and second, the effects of that reinforcement only propagate back towards the start state labelled "S" one level at a time. At a very minimum then the influence of the reinforcing goal state cannot reach the starting point until the animat has made many "forward" transitions. It might be conjectured that there is a form of "two-steps-back/one-step-forward" strategy that would optimally spread the goal's influence, but this would be a highly artificed strategy. In practice sufficient numbers of propagating transitions are not made until a large number of trials have been completed. Protracted learning rates are recognised as a limitation of this class of reinforcement learning algorithm (e.g., Wyatt, 1995). The protracted learning rate of this class of reinforcement algorithm provides an advantage in terms of noise immunity. The lack of immediate commitment allowing an accurate model of the variability to be constructed. SRS/E will be tested in a later experiment to determine the degree to which learning rate and task performance degrade under the noise conditions defined by Sutton.

Is it not the case then that all SRS/E is doing is recording every transition, building a simple graph and so easily traversing it? For Lprob = 1.0 the conditions for learning are indeed ideal under these experimental conditions. Each state is recognised by a unique and reliable identifier, every action reliably transitions between two such states, the μ-hypothesis creation mechanism explores exactly this relationship first, and the animat is permitted to learn *ad libitum*. Why should learning be anything other than *one-shot* when conditions are ideal? As these conditions move toward more realistic circumstances the expected, and observed, learning performance falls away from this ideal case. In doing so they repeatably demonstrate the forms of the learning curve so ubiquitously observed in experiments with animals.

Several reinforcement algorithms claim to achieve optimal performance over a fixed task of this nature[27], yet SRS/E does not demonstrate perfect performance even after 100 trials under the optimal conditions (Lprob = 1.0, figure 6-2). Recall that the average path length was 15.32 on the second trial, and improved only marginally to 14.96 after all 100 trials. Why should this be? SRS/E and reinforcement learning algorithms make fundamentally different assumptions. Dyna-PI is set a repetitive task and builds a static policy map. For every condition an optimal policy action is ultimately made available. By successively reducing the learning rates and action selection variability (by reducing the *Boltzmann distribution* "temperature") the policy map stabilises. Under these conditions it may be more germane to enquire how the performance of SRS/E improves at all while the goal is continually reasserted. The answer lies in the 16 run-on cycles following the animat's arrival at the goal location. Learning occurs independently of valenced behaviour and new μ-hypotheses can be created during this brief period.

SRS/E is specifically an algorithm for learning and behaviour. Goals arise, are satisfied (or not) and the animat moves on to some different activity. Once a goal is asserted the algorithm pursues it via the best path without additional exploration, using whatever information is available at the time. The experimental circumstances described here exclude any variability due to noise, so that when the goal is continually reasserted without interruption, the animat pursues the path without variation. Where an optimal path is located first, then all subsequent paths are also optimal, where a sub-optimal path is located, all subsequent paths will be sub-optimal. Under normal conditions the animat would pursue other activities, allowing new μ-hypotheses to be created, and so overall improvement in goal acquisition would occur over time. There is a detectable correlation between the amount of exploration during the random walk exploratory phase and the resulting average path length under valenced test conditions. Enabling the oscill (γ^4) component would explicitly add the dimension of exploratory behaviour, but would always tend to detract from the performance of optimal solutions.

[27] Notably those which reduce to an established *dynamic programming* technique and are thus able to exploit the existence of optimal solution proofs (Ross, 1983).

6.3. The Effects of Noise

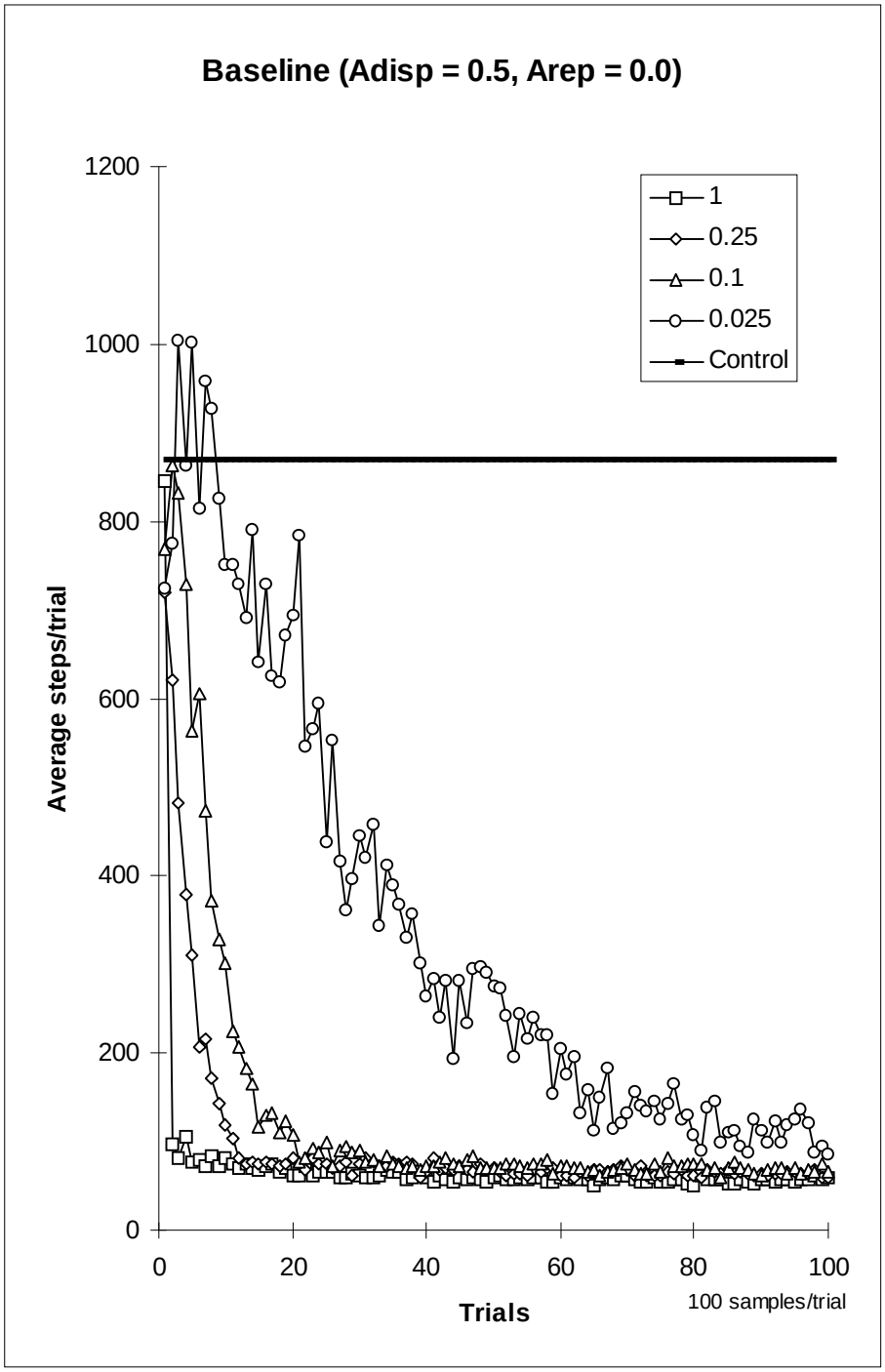
Sutton (1990) defines a test procedure for determining the effects of noise on the Dyna family of reinforcement algorithms. Noise, by Sutton's definition, perturbs the proper action of the animat by altering the effect of its actions, effectively after the animat has issued them, and so is completely outside the control of the animat. Provision for adding this form of noise is made within the SRS/E system. It is controlled by the action dispersion probability (Adisp) parameter. Adisp is selected by the investigator at the start of each experimental run. Its use and effects were described earlier in chapter five. This series of experiments is designed to evaluate the effects on both learning and valenced behaviour in SRS/E. Sutton did not publish noise results for the Dyna algorithms.

6.3.1. Description of Procedure

The experimental procedure described for the baseline experiments was repeated, with the exception that Adisp was set to 0.5 (50% of actions changed, 50% unchanged). The data from the total of 42,500 trials was recorded and plotted as before. A separate control line was determined for these experiments. The complete experimental procedure was then repeated with Adisp set to 0.75 (75% of actions unchanged, 25% changed).

6.3.2. Results and Analysis of Experiment

Figure 6-4 summarises the results from this investigation for Adisp = 0.5. Two points are of note. First is that the slope of the learning curve is not noticeably different for the results obtained in the noise free situation. Second the average valenced path length following stabilisation (as measured by the mean of the last 25 trials for Lprob = 1.0, 0.25 and 0.1, total of 7,500 individual trials) is markedly higher at 65.84 than that for the noise free case, at 15.46. There is also more variability in the valenced path lengths (as determined by the standard deviation, 45.99 as opposed to 1.34 for the noise free case). The Adisp = 0.75 trials resulted in a mean of 25.19 and a standard deviation of 14.42 under the same conditions. The learning curves in this case also showed a similar slope to the Adisp = 1.0 and 0.5 investigations.



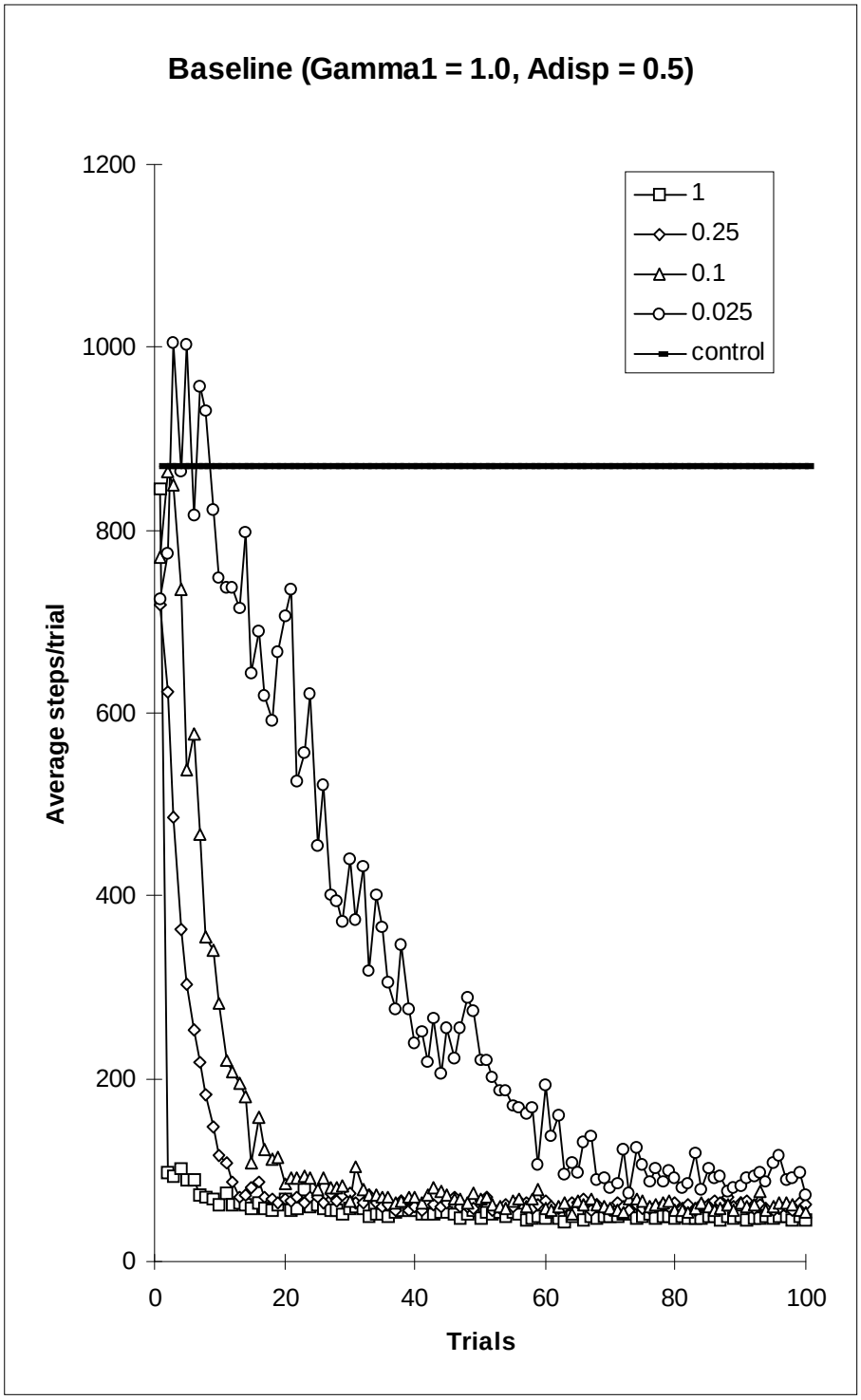
monolith\results\bse50all\bse50all.xls

Figure 6-4: Baseline Learning with Noise (Adisp = 0.5, Lprob = 1.0, 0.25, 0.1 and 0.025)

6.3.2.1.Tuning Parameters for Static Environments

The "standard" set of *selection factor* values ($\gamma^1 = 0.0$, $\gamma^2 = 0.9$, $\gamma^3 = 0.1$ and $\gamma^4 = 0.0$) was employed for the above investigations. These settings are appropriate to a changing environment, as the cost estimate values are biased toward more recent

events. The experimental environment used here is essentially static, apart from the introduced noise, the level of which remains constant. The investigation with Adisp = 0.5 was repeated (for Lprob = 1.0, 0.25, 0.1 and 0.025 over 100 runs each of 100 trials), with the value of γ^1 set to 1.0 (so $\gamma^2 = \gamma^3 = \gamma^4 = 0.0$). Cost estimates are therefore directly related to the probability of successful prediction of each μ-hypothesis. The estimates are calculated from the unadjusted count of frequencies of satisfied expectations to total activations from the cycle on which the μ-hypothesis was created. Figure 6-5 shows the resulting learning curves. Conditions were identical to the results shown in figure 6-4, except as indicated.



monolith\results\gamma1\base_g1.xls

Figure 6-5: Baseline with Noise (Adisp = 0.5, γ^1 = 1.0)

The average valenced path length following stabilisation (as measured by the mean of the last 25 trials for Lprob = 1.0, 0.25 and 0.1, a total of 7,500 individual trials) is indeed lower, at 56.83 (stddev = 56.18), than for the γ^2 = 0.9 case, (65.84 steps/trial), but still higher than that for the noise free case (15.46 steps/trial).

These results indicate that alterations in the cost estimation parameters have some effect, but that this is not as pronounced as might have been expected under these conditions.

6.3.2.2.The Effects of Noise: Learning or Behaviour?

The question remains whether the decrease in animat goal seeking performance is primarily due to inaccuracies in the Dynamic Policy Map, or a consequence of the disruption due to the animat's individual action selections being thwarted by the noise process. This detailed investigation takes a specific individual and allows it to run for 100 trials with the noise parameter Adisp set to 0.5 (to replicate the baseline run). The investigator then regains manual control of the experiment and forces the value of Adisp to 1.0 (no dispersive noise), returns the animat to the start location, enables the standard goal and records the number of steps taken. Figure 6-6 compares the two subsequent trial paths, trial 101 with Adisp = 0.5, and trial 102 with Adisp = 1.0.

(A) Path with Adisp = 0.5 (trial 101)

```
Cycle 7299: World is 6 by 9
   0    0    0    0    0    1    1 ****    1X
   0    0 ****    0    1    5    3 ****    1
   2    1 ****    0    1    3    2 ****    3
   1    0 ****    1    1    0    1    3    3
   1    3    1    1    0 ****    0    0    0
   0    2    0    0    0    0    0    0    0
Mean = 0.914894, std dev = 1.185013
```

(A) Path with Adisp = 1.0 (trial 102)

```
Cycle 7314: World is 6 by 9
   0    0    0    0    0    0    0 ****    1X
   0    0 ****    0    0    0    0 ****    1
   1    1 ****    0    0    0    0 ****    1
   0    1 ****    1    1    1    1    1    1
   0    1    1    1    0 ****    0    0    0
   0    0    0    0    0    0    0    0    0
Mean = 0.319149, std dev = 0.483779
```

monolith\figures.ppt:slide 4

Figure 6-6: a) Path with Adisp = 0.5 (trial 101), b) Adisp = 1.0 (trial 102)

Inspection of the Valenced Path printout (figure 6-8) from the experiment trace log file confirms the soundness of the valenced path created under noise conditions. Figure 6-7 shows the policy map generated at the conclusion of trial 101. Each location shows the appropriate action except X=5, Y = 0 (bottom row, fourth back from right corner). μ-Hypothesis H223 ("S28<X5Y0> → D → S29<X6Y0>") has an estimated cost of 3.0, 14 of the 42 activations to date having succeeded. The "correct" μ-hypothesis, H121 ("S28<X5Y0> → R → S29<X6Y0>") has an estimated cost of 4.66, only three of the 14 trials to date having succeeded. Such is the consequence of probabilistic dispersive noise. Each action is selected independently, there is no guarantee at any point the ratio of the three possible actions reflects the 0.5:0.25:0.25 selection process. The location is away from the

valenced goal path and consequently these policy recommendations were developed during the exploration period. Were this location to fall on the valenced path the system would naturally select H223. On the assumption it would fail in 75% of cases its estimated cost would eventually rise above that of H121, which would then become the preferred choice. Note that the majority of other estimated costs (line four in each location cell) more closely reflect the expected value of 2.0. Figures 6-6, 6-7 and 6-8 were all extracted from the latter ($\gamma^1 = 1.0$) investigation.

```
Policy map at cycle 7299
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|H164@14 |H378@13 |H29@12  |H45@11  |H276@10 |H380@9  |H148@8  |........|        |
|R       |R       |R       |R       |D       |D       |D       |........| GOAL   |
|   28.44|   26.64|   24.41|   22.28|   20.42|   18.68|   16.38|........|        |
|    1.80|    2.23|    2.13|    1.87|    2.00|    2.00|    1.70|........|        |
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|H1@15   |H18@14  |........|H109@10 |H48@9   |H49@8   |H301@7  |........|H493@1  |
|D       |D       |........|R       |R       |D       |D       |........|U       |
|   29.91|   28.20|........|   20.96|   18.42|   16.68|   14.68|........|    2.33|
|    1.38|    1.62|........|    2.55|    1.74|    2.33|    2.18|........|    2.33|
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|H70@14  |H176@13 |........|H305@9  |H213@8  |H50@7   |H331@6  |........|H492@2  |
|R       |D       |........|R       |R       |R       |D       |........|U       |
|   28.53|   26.58|........|   18.70|   16.52|   14.35|   12.50|........|    4.52|
|    1.94|    2.00|........|    2.18|    2.17|    1.85|    1.99|........|    2.19|
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|H192@13 |H20@12  |........|H100@8  |H294@7  |H382@6  |H383@5  |H384@4  |H422@3  |
|D       |D       |........|R       |R       |R       |R       |R       |U       |
|   26.66|   24.58|........|   16.26|   14.45|   12.67|   10.51|    8.51|    6.59|
|    2.03|    2.17|........|    1.81|    1.78|    2.15|    2.00|    1.93|    2.06|
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|H182@12 |H247@11 |H84@10  |H95@9   |H286@8  |........|H393@6  |H350@5  |H346@4  |
|R       |R       |R       |U       |U       |........|R       |U       |U       |
|   24.63|   22.41|   20.45|   18.22|   16.06|........|   12.33|   10.45|    8.63|
|    2.21|    1.96|    2.23|    1.96|    1.61|........|    1.88|    1.93|    2.04|
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
|H185@13 |H205@12 |H207@11 |H209@10 |H210@9  |H223@8  |H227@7  |H407@6  |H426@5  |
|R       |R       |R       |R       |U       |D       |R       |R       |U       |
|   25.43|   23.38|   21.51|   19.66|   17.85|   17.00|   14.00|   11.65|   10.10|
|    2.06|    1.87|    1.85|    1.81|    1.79|    3.00|    2.35|    1.54|    1.48|
+--------+--------+--------+--------+--------+--------+--------+--------+--------+
```

Figure 6-7: Policy Map at Conclusion of Trial 101

Separate observations from a number of individual runs from both investigations, and from inspection of Dynamic Policy Maps ("M" command) confirm that the effects on valenced path length are mainly from the execution of the behaviour, rather than faults in the μ-hypothesis creation process or construction of the policy map. "Inappropriate" actions still appear in the DPM, and may do so at any point in the investigation due to the chance of long sequences of noise affected actions altering the relative strength of the μ-hypotheses relevant to the achievement of any given location in the path. Clearly this is more likely in the case where learning is biased towards recent events. In this instance a long sequence of noise affected actions will have a disproportionate effect at any point in the animat's existence. Where $\gamma^1 = 1.0$ the same sequence of noise affected actions will have greater effect

while the total activations of the affected μ-hypothesis is low. In practice the system has shown itself (over thousands of trials) to be particularly tolerant of these chance events, re-establishing appropriate paths once the sequence of anomalous events is ended.

```
VBP @ 7256 = 285.322, bestcost = 28.5192
GOAL 46, Max valence level is 16
H70 predicts S5[X1Y3] from S0[X0Y3] (*active) after R (cost = 1.942029, total = 28.519203)
H176 predicts S6[X1Y2] from S5[X1Y3] after D (cost = 1.978261, total = 26.577173)
H20 predicts S7[X1Y1] from S6[X1Y2] after D (cost = 2.169492, total = 24.598913)
H247 predicts S22[X2Y1] from S7[X1Y1] after R (cost = 1.942308, total = 22.429422)
H84 predicts S23[X3Y1] from S22[X2Y1] after R (cost = 2.246154, total = 20.487114)
H95 predicts S26[X3Y2] from S23[X3Y1] after U (cost = 1.981482, total = 18.240959)
H100 predicts S20[X4Y2] from S26[X3Y2] after R (cost = 1.833333, total = 16.259478)
H294 predicts S25[X5Y2] from S20[X4Y2] after R (cost = 1.764706, total = 14.426144)
H382 predicts S33[X6Y2] from S25[X5Y2] after R (cost = 2.152542, total = 12.661438)
H383 predicts S40[X7Y2] from S33[X6Y2] after R (cost = 2.012987, total = 10.508896)
H384 predicts S42[X8Y2] from S40[X7Y2] after R (cost = 1.934307, total = 8.495909)
H422 predicts S44[X8Y3] from S42[X8Y2] after U (cost = 2.051020, total = 6.561603)
H492 predicts S45[X8Y4] from S44[X8Y3] after U (cost = 2.185185, total = 4.510582)
H493 predicts S46[X8Y5] (goal) from S45[X8Y4] after U (cost = 2.325397, total = 2.325397)
Valenced path in 14 steps, estimated cost 28.519203
```

Figure 6-8: Planned Valenced Path (trial 101)

6.3.3. Discussion

The introduction of dispersive noise into the SRS/E system is undoubtedly reflected in the performance of the animat under these controlled experimental conditions. These investigations also confirm that the learned component of the system is resilient to this form of noise (as is also claimed for certain *Q*-learning systems), actions derived from available μ-hypotheses at each choice point reflecting probabilities from past experience. The system may be made more or less reactive to change in the environment by the selection of parameters. Sutton (1990) suggests the possibility that a second order learning phenomena might be employed to determine the long term applicability to an individual animat of a particular strategy. Alternatively selection pressures within a population of individuals might be considered an appropriate strategy.

Dispersive noise, of the form investigated here is only one form of noise. The current implementation of SRS/E also allows for the introduction of random tokens into the input token stream. Such tokens emulate the presence of extraneous events, unrelated to the performance of the task. Using the postulate system described SRS/E will incorporate these random occurrences into μ-hypotheses as a matter of course. SRS/E will be sensitive to this form of noise. First in that it will

precipitate the formation of spurious μ-hypotheses, diluting the Hypothesis List and adding computational overhead. Second in selecting whatever response was incorporated into the spurious μ-hypothesis at the time of its creation, inappropriate actions will be selected in pursuit of the current top-goal. As the availability of more effective μ-hypotheses increases, these spurious μ-hypotheses will contribute less to the behaviour of the animat and will eventually be expunged by the μ-hypothesis deletion procedures considered in chapter four.

6.4. Alternative and Multiple Goals

These investigations demonstrate the effect of the SRS/E system when confronted with several different goals, either sequentially or simultaneously. The results of these investigations illustrate the manner in which SRS/E handles goals and valenced behaviour, and highlights the differences between the Dynamic Expectancy Model and reinforcement learning methods that create a static policy map.

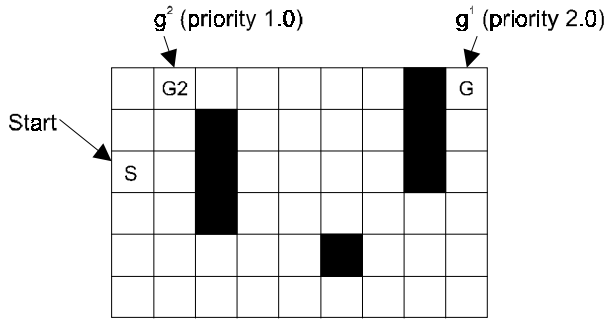
6.4.1. Description of Procedure

In investigation one of this experiment naïve animats are allowed an exploration period in the chosen environment, in this instance DynaWorld/Standard (figure 5-1). Each run uses the defined starting point ("S"). The initial unvalenced trial-and-error exploration period is chosen to allow the animat adequate opportunity to thoroughly explore its environment (1,000 execution cycles). An action repetition rate (Arep) value of 0.5 is selected to reduce initial random-walk time. The unvalenced time to reach the goal is noted. At the end of the exploration period the animat is returned to the known starting point, and the goal state ("G") is asserted with a priority of 1.0. The valenced time to reach to Goal is noted. On reaching the standard goal ("G") the original starting location ("S") is now asserted as the goal, with a priority of 1.0, and the valenced time for the animat to re-traverse the environment noted. To confirm these findings these two traversals are repeated, and the respective valenced path times noted.

As a control, investigation two of the alternating goal experiment repeats investigation one of the experiment with the start and goal locations reversed at

every stage in the procedure. The procedure is repeated 10 times and the results tabulated. A single instance is selected and individual paths presented for detailed discussion.

The third investigation of this experiment presents the animat subjects with two goals simultaneously. The path generated to reach these two goals should verify the mechanism by which SRS/E seeks and satisfies elements on the Goal List G. Individual naïve animats are given an identical training period to the previous investigations using the DynaWorld/Standard environment, before being returned to the start location "S". Two goals are then enabled simultaneously, one of which is the original goal ("G"), with a priority of 2.0, and the other chosen to be at some location ("G2" at X = 1, Y = 5) on or near an expected valenced path between start and original goal. The goal "G2" is assigned a lower priority (1.0), figure 6-9.



Graphic 5.12 from monolith\mazes.cdr

Figure 6-9: Simultaneous Goal Locations

6.4.2. Results and Analysis of Experiment

Results for the first investigation are shown in table 6-1. The first column indicates the starting random seed, the second the number of actions taken during the random walk to reach the location "G". The goal is not asserted and so has no special significance to the animat at this stage. The third shows the length of the valenced path for the first traversal from "S" to "G". The fourth column records the length of the valenced path returning from "G" (as starting point) to "S" (now valenced as the goal). The fifth and sixth columns record the valenced path lengths from "S" to "G" (valenced) and then from "G" to "S" (valenced) respectively. The animat position is only changed by the investigator once, directly following the random-walk period.

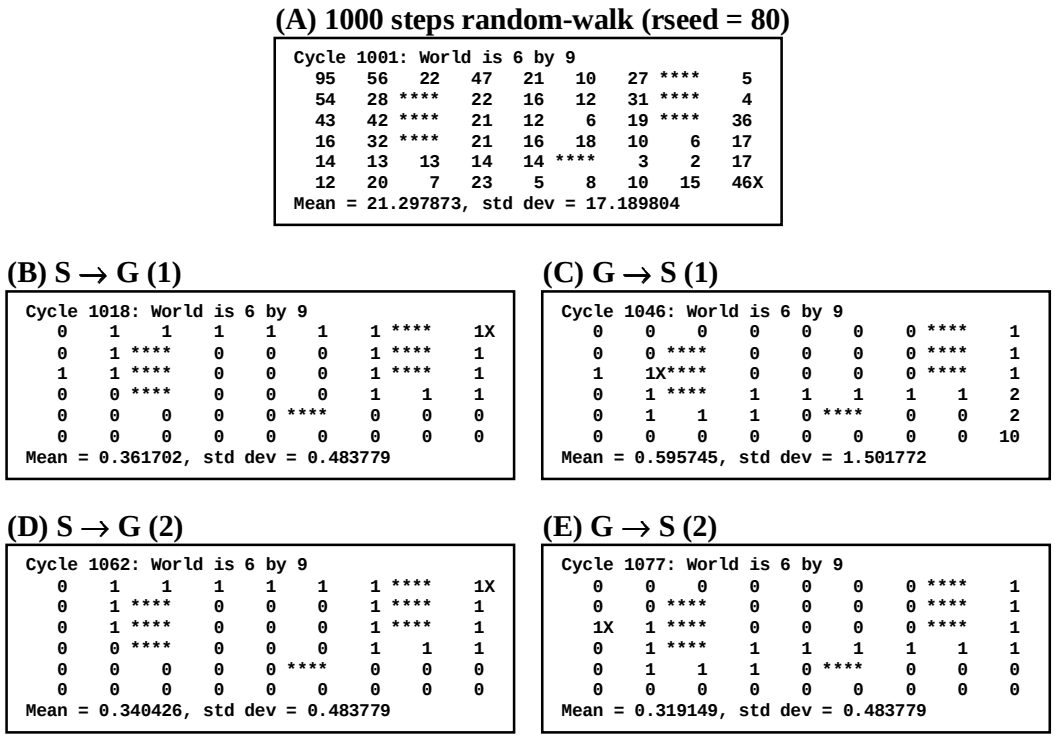
Under these essentially ideal learning conditions the initial valenced path from "S" to "G" is close to the minimum. The variation observed is consistent with the observation that the 1000 random-walk cycles was insufficient to completely build the full potential Hypothesis List, so solution paths may be sub-optimal. The first return path ("G" → "S") consistently requires more cycles than would be expected following this level of experience. Figure 6-10 details the individual animat paths at different stages in a single experimental run and indicates the reason for the apparently anomalously extended path length. Figure 6-10a records (shown using the "W" command) the number of visits by the animat to each location during the exploratory, unvalenced, random-walk period. The location cell labelled "X" (X=8, Y=0) indicates the position of the animat when it was removed by the investigator to the start location for the first valenced run. Figure 6-10b shows the first valenced path, non-optimal at 16 steps, no doubt as a consequence of the greater degree of exploration in the upper part of the environment on this particular run.

Seed	1st visit "G"	S→G (1)	G→S (1)	S→G (2)	G→S (2)
10	915	16	23	15	14
20	317	14	28	13	14
30	216	14	18	13	15
40	101	15	15	13	16
50	534	14	19	15	14
60	167	14	14	15	13
70	379	14	18	15	16
80	265	16	27	15	14
90	134	14	33	15	16
100	140	14	29	13	14
Average	316.8	14.5	22.4	14.4	14.6

Table 6-1: Results for Investigation One of Dual Goal Experiment

Figure 6-10c shows the return path. The animat moves to location (X=8, Y=0) immediately and appears to become trapped there for some number of execution cycles, thereby increasing the overall path length to 27 (from a possible 14). This is

an experimental artefact, demonstrating that this emulation of learning and behaviour requires as much care in the conduct of experimental procedure as does work with real animal subjects. The forcible movement of the animat to the start location caused a spurious μ-hypothesis ("H167: <X8Y0> → D → <X0Y3>")[28] to be created, which promises a short-cut to the current goal location. The μ-hypothesis H167 fails to deliver this promise at every trail. Its cost estimate contribution increases at each attempt until it exceeds that for the effective path, which is adopted at the next DPM rebuild. When this path is again valenced, the shorter path is adopted immediately, figure 6-10e.



monolith\figures.ppt:slide 5

Figure 6-10: Animat Random and Valenced Paths (investigation 1, rseed = 80)

Table 6-2 records the results of investigation two of this experiment, where the roles of "S" and "G" from figure 6-1 are reversed throughout the procedure. The results are broadly similar to those of investigation one and clearly demonstrate that these results are independent of the actual start and goal locations.

[28]"H167 predicts S0[X0Y3] (goal) from S36[X8Y0] after D (cost = 1.818182, total = 1.818182)": from the valenced path summary recorded in the experiment trace file.

Seed	1st visit "S"	G→S (1)	S→G (1)	G→S (2)	S→G (2)
10	125	16	33	16	15
20	113	14	28	14	13
30	355	16	22	15	13
40	355	16	24	15	15
50	103	16	29	16	13
60	228	14	35	14	15
70	921	16	15	16	15
80	111	14	15	14	15
90	66	14	18	14	15
100	216	14	17	13	13
Average	259.3	15.0	23.6	14.7	14.2

Table 6-2: Results for Investigation Two of Dual Goal Experiment

Table 6-3 summarises the results obtained for the simultaneous goal procedures of investigation three. The effect of setting these two goals is to cause the animat to visit each in turn. In the majority of cases the animat visits the more distant, but higher priority goal first, and then doubles back to satisfy the secondary lower priority goal. The average valenced path length to the first goal is 14.33, and the average total travel to both goals is 32.44. The disruptive effects of the forced return to "S" are still apparent. In one instance the goals are visited in the reverse order (rseed = 80), with valenced path lengths of 3 and 16 respectively. This is purely because the secondary goal lay on the path taken by the animat to the primary goal. A goal is satisfied by being achieved, regardless of whether or not this was because of a valenced action specifically intended to satisfy that goal. The use of "cloned" animats for parts 1 and 3 of this experiment means the initial exploratory and first goal paths are identical.

Seed	1st Visit "G"	1st Goal	2nd Goal
10	915	16	29
20	317	14	39
30	216	14	27
40	101	15	32
50	534	14	27
60	167	14	27
70	379	14	35
80	265	3	16
90	134	14	37
100	140	14	39
Average	316.8	13.2	30.8

Table 6-3: Results for Investigation Three, Simultaneous Goals

Figure 6-11 shows two individual goal paths. Figure 6-11a records the path for rseed = 30, and is typical of the situation where the primary goal is visited first, then the secondary goal. Figure 6-11b shows the situation where the secondary goal is satisfied first because it happens to lie on the valenced path to the primary goal (rseed = 80).

(A) S → G1 → G2 (14/27 steps, seed = 30)

```
Cycle 1029: World is 6 by 9
   0    1    1    1    1    0    0 ****    1
   0    0X****    0    1    1    1 ****    2
   1    0 ****    0    0    0    1 ****    2
   1    0 ****    1    1    1    2    2    2
   1    1    1    1    0 ****    0    0    0
   0    0    0    0    0    0    0    0    0
Mean = 0.595745, std dev = 0.684167
```

(B) S → G2 → G1 (3/16 steps, seed = 80)

```
Cycle 1018: World is 6 by 9
   0    1    1    1    1    1    1 ****    1X
   0    1 ****    0    0    0    1 ****    1
   1    1 ****    0    0    0    1 ****    1
   0    0 ****    0    0    0    1    1    1
   0    0    0    0    0 ****    0    0    0
   0    0    0    0    0    0    0    0    0
Mean = 0.361702, std dev = 0.483779
```

monolith\figures.ppt:slide 5

Figure 6-11: Sample Simultaneous Goal Paths

6.4.3. Discussion

These investigations show substantial differences between existing reinforcement learning methods and the SRS/E algorithm. Goals may be selected at will from the available elements in the Sign List, and a Dynamic Policy Map built from the

available μ-hypotheses to attempt a solution path. A standard reinforcement or *Q*-learning algorithm would presumably have to completely rearrange the static policy map over many trials before reasonable performance to the new goal is re-established. As reinforcement does not take place until the changed goal is achieved, if that new goal did not fall on the solution path to the previous goal, this might never happen. This result from the Dynamic Expectancy Model is considered a significant challenge to conventional reinforcement learning algorithms.

Investigation three of this experiment demonstrates SRS/E's flexibility and effectiveness in handling multiple goals. Much progress has been made in adapting reinforcement algorithms to build several policy maps to address multiple goals (section 2.4.2). This approach brings a severe computational cost penalty as the number of recorded goals increases, and means that all goals must be identified before learning can take place. These limitations do not apply to SRS/E. Section 7.2 proposes some extensions to SRS/E to modify its goal seeking behaviour to balance the estimated cost of achieving a goal with the given priority of the goal.

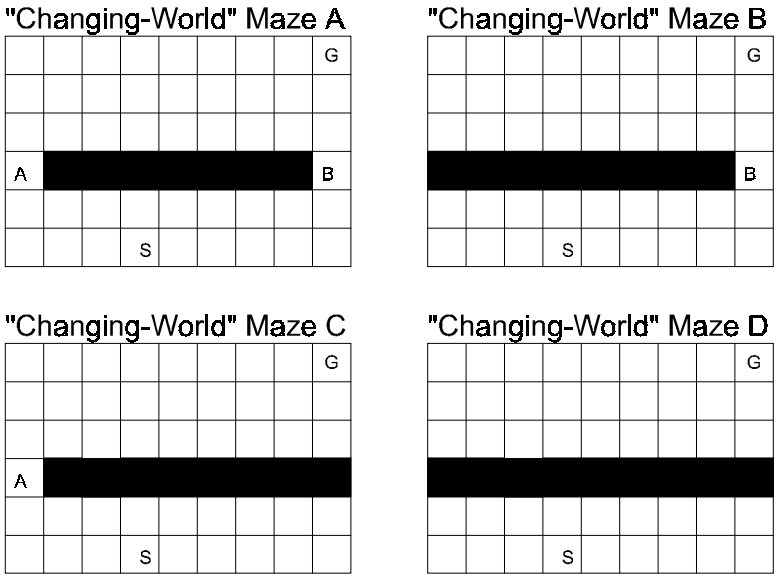
6.5. Multiple-Path, Blocking, Shortcut and Extinction Investigations

The individual investigations in this experiment series evaluate the performance of SRS/E in a range of conditions where multiple paths exist, become available, or cease to be available, between a constant start and constant goal location. The first investigation determines the learned behaviour of SRS/E in an environment where two distinct paths, one longer than the other, exist between start and goal (*multiple-path*). The animat has been allowed to adequately explore the environment fully before the start of the investigation. The investigation further determines the effect of blocking the preferred route.

In the second investigation the effects of blocking one previously explored and known path, and then two known paths is considered. This investigates the *extinction* phenomena, where a goal is abandoned as unattainable. The third investigation repeats a procedure reported by Sutton (1990) to determine the enhanced performance of his Dyna-Q+ system, compared with Dyna-PI, when presented with the situation where a known short path becomes blocked, and a previously unknown path is released (*path blocking*). Results of this latter

investigation are presented in a manner comparable to that employed by Sutton. Finally the performance of SRS/E and programs from the Dyna family are considered in a situation where a previously unknown shortcut is introduced.

This series of investigations uses an experimental environment described by Sutton (1990) and shown in figure 6-12. Start “S” and Goal “G” locations are the same throughout the investigation. Obstructions are selectively added or removed during individual investigations at the points marked “A” and “B”.



Graphic 5.15 from monolith\mazes.cdr

Figure 6-12: Changing World Environments

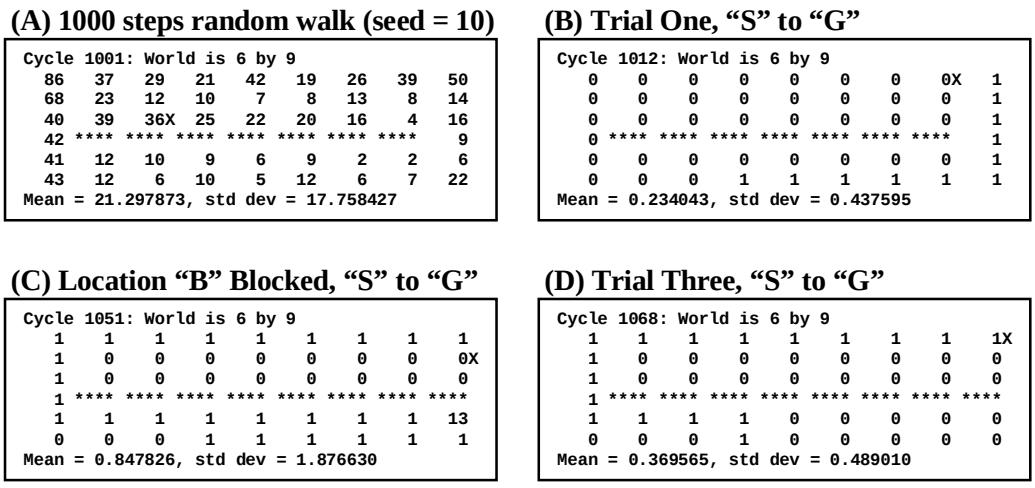
6.5.1. Investigation One (Multiple-Path), Procedure

This investigation determines the actions of an animat in an environment with two known paths, one of which is shorter than the other. Under these circumstances the animat is expected to take the shorter of the paths (that of lower estimated policy cost), but select the longer path should the shorter become unavailable. In this investigation the animat is allowed to explore the environment of figure 6-12a for 1000 cycles as a random walk with no goal asserted. With Arep is set to 0.5, this allows sufficient time for the environment to be completely explored. On completion of this first phase the animat is returned to “S” and goal “G” asserted with a priority of 1.0. The investigator confirms that the animat reaches the goal by the shorter of the alternative routes (i.e., via location “A”). The number of steps is

noted. The animat is returned to “S” and location “B” is blocked. Goal “G” is again asserted with a priority of 1.0 and the behaviour of the animat noted. The animat is returned to “S”, “G” asserted and the resulting path noted.

6.5.2. Investigation One, Results and Analysis

Figure 6-13 shows the effect on animat behaviour of the procedure described for investigation one. The 1000 cycles of random walk provide ample opportunity for the animat to discover both available paths (figure 6-13a). Figures 6-13b, c and d show the animat path from “S” to “G” with no additional obstruction, the first run after location “B” is obstructed and the second run after “B” is obstructed respectively. This investigation was repeated with ten individual animats (rseed = 10, 20 .. 90, 100), the instance shown is with individual rseed = 10. With no dispersive noise and Lprob = 1.0 performance across these individuals is constant, the average first path length being 10 steps, and the third 16 steps. The average second path length is 39.7. Nine of the individuals took 39 steps. One 46 due to the appearance of a spurious shorter route μ-hypothesis introduced by handling during the procedure (the forced return move to “S” fell, by chance, in the lower right catchment area).



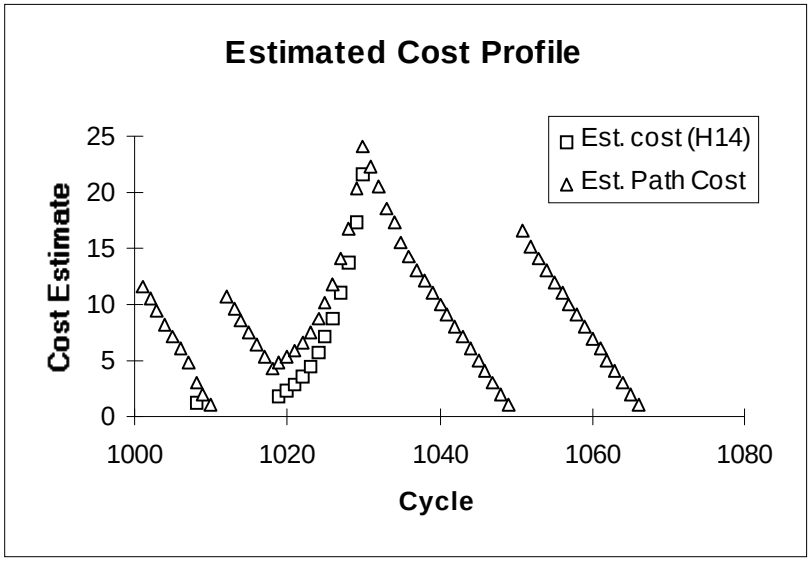
monolith\figures.ppt:slide 7

Figure 6-13: Multiple Path Investigation, Individual rseed = 10

The mechanism by which SRS/E selects the original path, and then selects and stabilises on the new path after the obstruction is detected is straightforward. The first path is the lowest cost path computed by the Dynamic Policy Map from

elements in the Hypothesis List. On the second trial run the DPM indicates the same path as run one. On reaching location X=8, Y=1 the previously reliable action "U" (from H14) fails, and the estimated cost of the step increases. The animat repeats this action until the estimated cost of the failed step raises the total estimated path cost above that for the alternative known route via location X=0, Y=2 (in the exemplar instance, 20.27). At this point the DPM is recomputed with the new shortest route and the animat pursues the new route to the goal.

Figure 6-14 details the cost estimate profile of the three valenced paths for the selected individual. The overall estimate for the remaining path is shown with triangle markers. The first series (cycles 1001 to 1011) shows the uninterrupted path from "S" to "G" via location "B". The second series (cycles 1012 to 1051) starts similarly to series one until the blocked location is detected. Estimated path cost increases as the cost contribution of the failed μ-hypothesis H14 increases (H14's contribution to the path cost is shown with square markers). Eventually the estimated cost of the preferred path exceeds that of the alternative, then the DPM policy estimates radically change and the animat follows the new path via location "A" without further interruption (cycles 1030 to 1051). The third series (cycles 1052 to 1067) confirms the preference for the new, longer, path.



monolith\results\chngwld\p14.xls

Figure 6-14: Estimated Cost Profile (Path and H14)

The apparent persistence with which the animat pursues the newly failed μ-hypothesis (H14) is determined primarily by the *extinction rate*, β. Within a normal population of individuals one might expect a range of values for this parameter and so the number of failed attempts to vary between individuals before the alternative path is adopted. The animat should not necessarily abandon its attempts at a known path too soon, as there are many circumstances where continued attempts are indeed better than not doing so. Mott's *ALP* robot controller being a case in point, the degree of persistence in goal seeking inadequately reflecting the rarity of the events sought. Other strategies could be proposed, including relating the degree of persistence rate to the existing quality and maturity of the μ-hypothesis in question.

6.5.3. Investigation One, Discussion

The ability of an animat to select an alternate, known, route if thwarted in pursuit of its preferred solution may appear as seemingly trivial. Yet this ability is an important discriminator between pure reinforcement learning systems and sensory-motor and intermediate level cognitive systems. Reinforcement learning systems (such as Dyna) which build a static policy map based on a current sensory pattern would not be expected to demonstrate the clear shift of behaviour presented by SRS/E, based as it is on a Dynamic Policy Map. Mimicking this ability therefore remains a challenge to conventional reinforcement learning systems. The distinction arises from the difference between categorising situations relative to a stable, but distant, reward and the encapsulation of situation and response as an independent unit disassociated from external reward.

6.5.4. Investigation Two (Goal Extinction), Procedure

This investigation determines the *goal extinction* behaviour of the animat when a single, known, path to the goal is obstructed, so that there is then no path to the goal. The animat is allowed to explore the environment shown in figure 6-12b for 1000 cycles (other conditions as for investigation one). The animat is returned to "S" and the goal location "G" asserted with a priority of 1.0. The animat's path to the goal noted. The animat is returned to "S", the location "B" blocked (so that there is no possible route to the goal) and goal "G" reasserted with priority 1.0. The behaviour of the animat in pursuing this unattainable goal is noted. The

investigation is repeated with the initial conditions from investigation one (figure 6-12a), where there are two initially available paths, with both paths being blocked at the end of the period of random walk exploration. The behaviour of the animat is noted under these conditions.

6.5.5. Investigation Two, Analysis of Results

Figure 6-15 shows the stages in the goal extinction process. Sub-figures 6-15a and b show the initial stages for this investigation (for the individual rseed = 10), the random walk exploration and the demonstration of successful valenced goal seeking behaviour given an unblocked path. The path to the goal is blocked at this step, the animat returned to "S" and the goal "G" reasserted. Sub-figures 6-15c to h show the stages in the extinction process. Initially valenced goal seeking behaviour proceeds as normal. As there is no alternative path the animat repeats the failed μ-hypothesis (H14) until the estimated cost of the path exceeds that for the *valence break point* (VBP) value calculated from the original cost estimate (10.28) for the path. At this point the animat reverts to unvalenced behaviour for a period regulated by the *goal recovery mechanism*, figure 6-15d. This period of exploration allows the animat to discover some new and previously unknown path to the goal (it would have already tried other possible paths had they previously been identified during the exploration phase).

(A) 1000 steps random walk (seed = 10)

```
Cycle 1001: World is 6 by 9
  93   27   31   17   36   20   38   19   44
  52    9   13    7    5    8   14    7   15
  79   17   16    8    4    5   14   13   18
**** **** **** **** **** **** **** ****   32
  43   11   13   11   18   15X   8    8   11
  40   16   25   25   17   29   16   13   21
Mean = 21.760870, std dev = 17.820969
```

(B) Test Valenced Path, "S" to "G"

```
Cycle 1012: World is 6 by 9
   0    0    0    0    0    0    0    0    1X
   0    0    0    0    0    0    0    0    1
   0    0    0    0    0    0    0    0    1
**** **** **** **** **** **** **** ****    1
   0    0    0    0    0    0    0    0    1
   0    0    0    1    1    1    1    1    1
Mean = 0.239130, std dev = 0.442326
```

(C) Valenced to Step 1039

```
Cycle 1039: World is 6 by 9
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
**** **** **** **** **** **** **** **** ****
   0    0    0    0    0    0    0    0   21X
   0    0    0    1    1    1    1    1    1
Mean = 0.600000, std dev = 3.094799
```

(D) Unvalenced to Step 1140

```
Cycle 1140: World is 6 by 9
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
**** **** **** **** **** **** **** **** ****
  22   16X   6    3    4    9    1    1    7
  12    3    6    0    0    2    2    2    5
Mean = 2.244444, std dev = 4.553387
```

(E) Valenced to Step 1159

```
Cycle 1159: World is 6 by 9
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
**** **** **** **** **** **** **** **** ****
   0    1    1    1    1    1    1    1   12X
   0    0    0    0    0    0    0    0    0
Mean = 0.422222, std dev = 1.782632
```

(F) Unvalenced to Step 1360

```
Cycle 1360: World is 6 by 9
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
**** **** **** **** **** **** **** **** ****
  45   18    6    3   12    5   16    9    6
  27   14    1    1    3    6   12   13   4X
Mean = 4.466667, std dev = 8.615232
```

(G) Valenced to Step 1371

```
Cycle 1371: World is 6 by 9
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
**** **** **** **** **** **** **** **** ****
   0    0    0    0    0    0    0    0   10X
   0    0    0    0    0    0    0    0    1
Mean = 0.244444, std dev = 1.483240
```

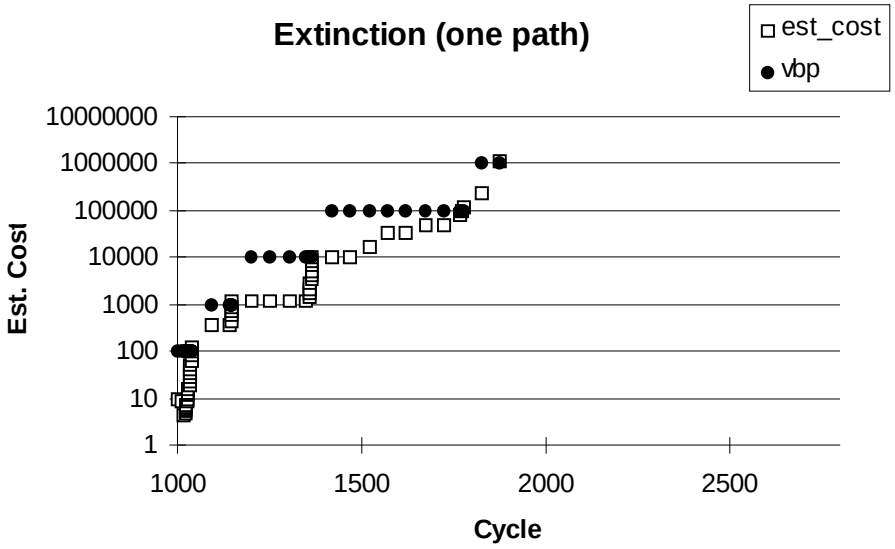
(H) Extinguished at Step 1593

```
Cycle 1593: World is 6 by 9
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0
**** **** **** **** **** **** **** **** ****
   0    4    6   10   10   12   11    9   43
   7    2    6   18   13X  11   11   23   26
Mean = 4.933333, std dev = 8.667949
```

monolith\figures.ppt:slide 8

Figure 6-15: Goal Extinction (rseed = 10)

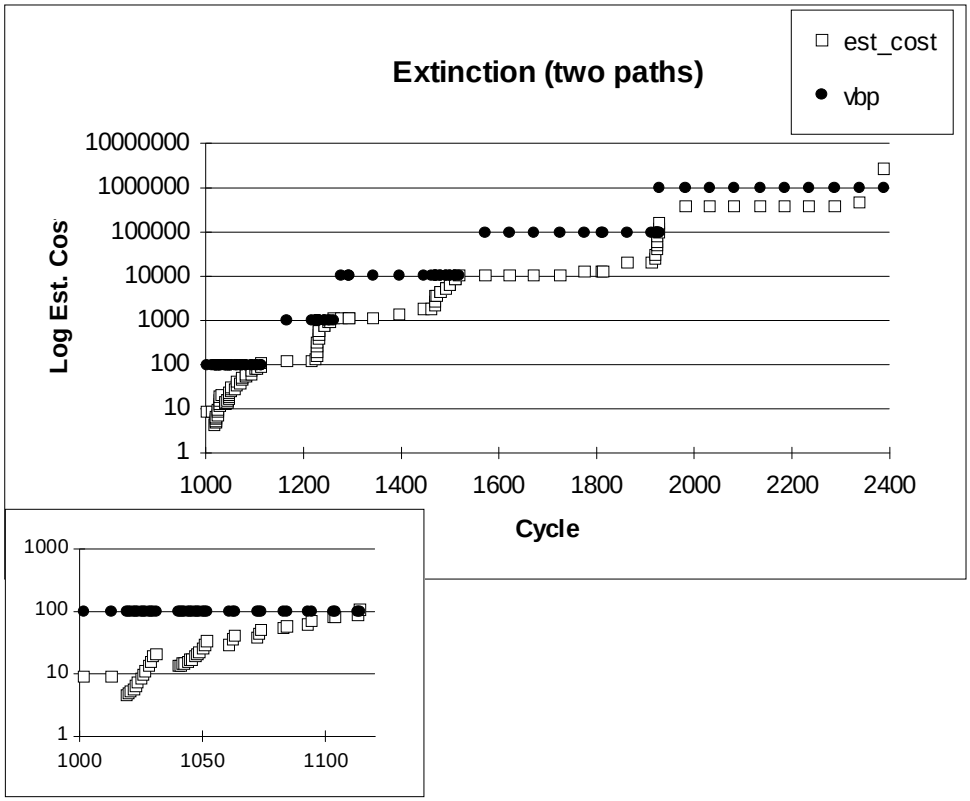
This process is repeated with alternating periods of valenced and unvalenced (trial and error) behaviour until the total cost estimate for the goal path exceeds the *goal cancellation level*, Ω, figure 6-15h. At this point g^1 is forcibly removed by SRS/E from the Goal List. The Innate Behaviour List $\mathcal{B}^g$ might reassert the goal, but to little useful effect. Figure 6-16 records the relative values of the cost estimate for the goal path and the computed value of VBP. Note in particular that the estimated cost rises quickly to meet the VBP at the end of each period of unvalenced behaviour. Note also that the estimated cost can rise during this unvalenced period due to the animat testing μ-hypothesis on the valenced path, but purely as a consequence of trial and error activities. This is particularly apparent in the latter stages of the extinction process and is in no small part due to the confined space in which the animat operates.



monolith\robtest\chngwld\extnct1g.xls

Figure 6-16: Goal Extinction, Comparison of Cost Estimate to VBP

This investigation was repeated with both paths ("A" and "B") available during the 1000 step random walk exploration phase (figure 6-12a). Both paths are then blocked before starting the extinction phase (as figure 6-12d). The animat behaviour is modified to appearing to scuttle back and forth between the two previously effective paths during the periods of valenced activity. Figure 6-17 shows the resulting estimated cost and VBP values of this investigation. The insert to the figure shows the detailed effect of this scuttling behaviour. Each rise in the cost estimate arises from the animat attempting the blocked μ-hypothesis, first at one end, and then at the other. The animat appears decreasingly persistent in its attempts to traverse each blocked path with each attempt. Gaps between the rises indicate the cycles during which the animat is (under valenced control) travelling between the two places where the known paths had been located. Note that the cost estimate and VBP are not shown during these periods as they are only recomputed when an event causes changes in Δ or δ that exceed `REBUILDPOLICYTRIP`. The net effect is to increase the number of cycles that elapse before goal extinction takes place. Over 10 separate trials (rseed = 10, 20 .. 100) the average time to extinction was 870.9 cycles for the single path case, and 1,443.2 cycles for this dual path case.



monolith\figures.ppt: slide 9 (monolith\robtest\chngwld\extnct2c.xls)

Figure 6-17: Goal Extinction (Two Path), Cost Estimate and VBP

6.5.6. Investigation Two, Discussion

Goal extinction phenomena are well documented for natural learning, and are supported by a wealth of experimental data. The rate at which extinction takes place appears to be highly variable. Razran (1971, p. 167) points out that under some operant conditioning regimes pigeons will continue with ineffective pecking behaviour (introduced with food reward) for over 10,000 events, expending more energy than would have been obtained from the reward. Classical conditioning regimes tend to demonstrate much more rapid extinction phenomena (Razran posits a median conditioning-extinction ratio of 36:1). The number of unrewarded actions required to produce goal extinction appears to depend on many factors including experimental conditions and procedures, the nature of the reward, its presentation and subject animal.

The onset of extinction can be continuously delayed by occasional reward (as in variable reward ratio regimes). Such is also the case in SRS/E where a single valid prediction restores the value of `bpos` for any μ-hypothesis disproportionately to the

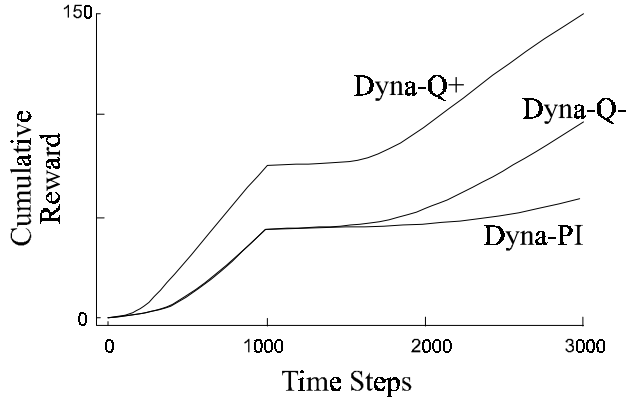
effect of a failed prediction. In its current implementation SRS/E does not demonstrate any spontaneous recovery of extinguished valenced behaviour. Such phenomena might be implemented by either an explicit second order term in the cost estimate function or by the inclusion of a specific *habituation* process disadvantaging μ-hypotheses used repeatedly. This would reflect Hull's approach to the extinction process (section 2.2, eqn. 2-1).

The presentation of data in figures 6-16 and 6-17 mirrors that for experimentally observed extinction patterns in animals (figure 3-1). Note that while these two presentations appear superficially similar they are not directly comparable, though they may indicate a similarity in underlying mechanism. The data in the figures presented in this chapter record internal values, those for animal experiments record externally observed events. Extinction in natural learning is a subtle phenomenon, no doubt deserving of a more sophisticated model that currently provided for in the SRS/E algorithm.

6.5.7. Investigation Three (Path Blocking), Procedure

This investigation determines the behaviour of an animat when faced with a block to a known path, but where a previously unknown path is simultaneously made available. To locate the new path the animat must balance exploration of the environment with exploitation of the previously known, and successful, solution path. In this investigation the animat is allowed a period of 1,000 cycles of continuously valenced activity using the maze shown in figure 6-12b (shorter path). The animat is always started at "S", with "G" asserted as goal. Once the animat reaches "G", it is returned to "S" and "G" reasserted. The other investigations in this experiment allowed random walk exploration during this initial phase. As in pervious experiments a small number of run-on cycles are permitted to ensure SRS/E may learn the steps leading directly to the goal. At cycle 1000 the location "B" is blocked and the previously blocked location "A" opened. The animat must discover the new path and continue to traverse from "S" to "G" as in the first phase of the investigation. Figure 6-18 shows the results obtained by Sutton (1990) for this blocking task with the Dyna family of reinforcement learning algorithms. The procedure used here follows that employed by Sutton. Effects of slight

variations in experimental procedure will be noted and discussed. The procedures for this investigation are available as a fixed schedule within SRS/E.



Graphic 5.21 from monolith\dyna.cdr

Figure 6-18: Average Performance of Dyna Systems on a Blocking Task

From Sutton (1990), p 222.

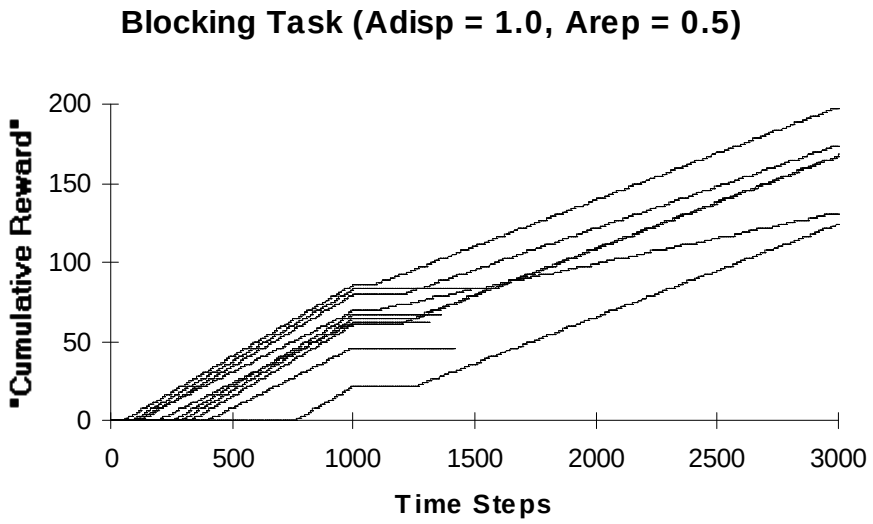
This investigation retains a cumulative record of the number of visits to the goal location, referred to as *cumulative reward* in figure 6-18. The slope of the line reflecting the frequency with which the goal is achieved. The shorter path allows the slope to be steeper, a flat period indicates a section in the investigation during which no "reward" is received, after location "B" is blocked and "A" opened. Results are plotted as curves recording individual animat performance and as an average of many individuals. Results for SRS/E are obtained with no dispersive noise (Adisp = 1.0), and with 10% dispersive noise (Adisp = 0.9).

6.5.8. Investigation Three, Results and Analysis

Figure 6-19 shows 10 individual performance curves for the conditions described by Sutton for the path blocking experiments (rseed = 10, 20 ... 100). As with figure 6-18 the slope of each curve indicates the path length from "S" to "G", the steeper the slope the more frequently the goal is visited. This form of presentation is analogous to that often used in *Skinner box* experiments to record the bar pressing activity of experimental animals in relation to reward delivery. Flat sections on a curve indicate periods where no reward is obtained. The first flat section indicates the initial random walk trial and error path to the goal. As Lprob is set to 1.0 in

these investigations the slope of the curve represents the length of the learned path (sometimes optimal, 7 cases of ten, sometimes not).

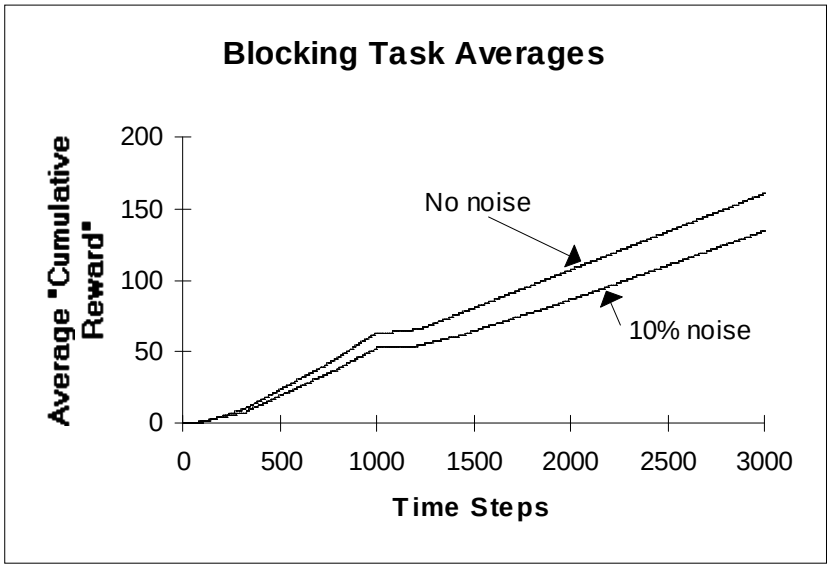
The second flat portion represents the time taken for the longer path to be located by trial and error random walk during the unvalenced parts of the goal extinction process. In four of the ten instances (individuals with rseed = 10, 50, 60 and 80) goal extinction took place before the alternative route was located. The cumulative curve ends abruptly in these cases. Members of the Dyna family of systems do not employ this mechanism. Of the remaining six individuals four found the shortest path from "S" to "G".



monolith\robtest\blocking\block1.xls

Figure 6-19: Investigation Three, Individual "Cumulative Reward" Curves

Figure 6-20 shows the averaged results of the ten individual trials described above. The performance of SRS/E under these conditions is comparable with the best of the Dyna series, *Dyna-Q+*, under similar experimental conditions (see discussion below). Addition of 10% dispersive noise (lower curve) has a consistently adverse effect on the performance of this system. The advantage of any additional exploratory effect being completely masked by the extra effort required to reach the goal. This finding appears consistent with previous conclusions about the effects of dispersive noise.



monolith\robtest\blocking\averages.xls

Figure 6-20: Investigation Three, Average "Cumulative Reward" Curves

6.5.9. Investigation Three, Discussion

Being fully aware of the difficulties of taking accurate measurements from a published graph (figure 6-18), a line drawn tangential to the first portion of the Dyna-Q+ curve indicates a slope of 10.76 steps/reward and for the second portion of the curve a slope of 18.2 steps/reward. Minimum path lengths are 10 and 16 respectively. Compensating for run-on cycles called for in the current experimental procedures, SRS/E attains average slope values of 10.6 and 18.33 respectively. It would be unreasonable to directly compare the total number of cumulative rewards at cycle 3000 (about 150 for Dyna-Q+, 160.33 for SRS/E) as the four worst instances in SRS/E were abandoned due to the extinction process. By adjusting the parameters involved SRS/E could be tailored to allow greater periods of random walk exploration during the unvalenced stages of the goal extinction process.

Sutton also tested members of the Dyna family of systems on a shortcut task. Animats were set a repeated goal seeking task using maze C (figure 6-12) in which only the longer path via "A" is available initially. After 3,000 cycles the shorter path "B" is also made available. Dyna-Q+, with its additional exploration component demonstrated some improvement in performance, indicating the shorter

path had been discovered and adopted. SRS/E has no explicit mechanism for exploration during valenced goal seeking behaviour. Consequently, if SRS/E is continuously tasked it will always adopt the best known path. Such wilful overtasking is a pathological case for SRS/E, the system expects to be presented with a range of tasks and to have periods where no goal is asserted. Under such conditions SRS/E has every opportunity to locate and subsequently employ the shortcut route.

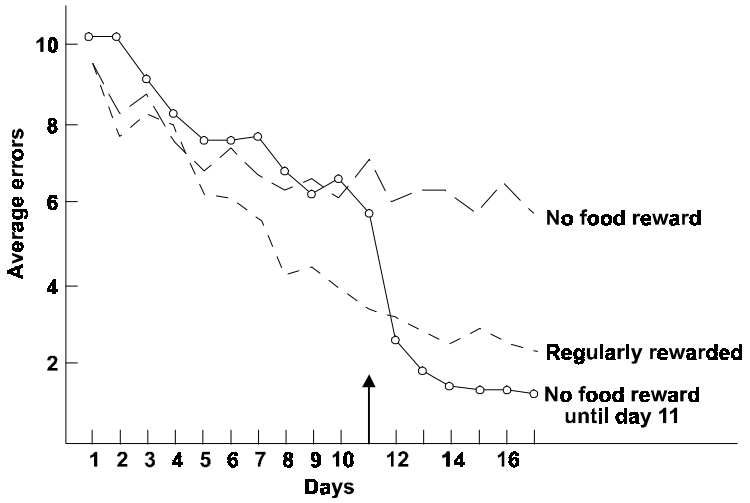
6.6. Latent Learning

The demonstration of *latent learning* phenomena was a significant step in the historical development of learning theory. Each of the major behaviourist learning theories is based on the notion that learning takes place in response to a reward (or conversely a punishment). If it were to be demonstrated that learning had occurred without any reward then the findings of the behaviourist school would be called into question. Clearly a demonstration of this type would have suited Tolman in the promotion of his expectancy theory.

A classic "latent learning" experiment is replicated with SRS/E. In the original Tolman and Honzik (1930) tested three groups of food deprived rats in a maze apparatus. The first group were allowed to wander the maze and obtained a food reward at the end location. The second group were allowed to wander the maze, but on reaching the end location they received no food reward. Each rat was placed in the maze once per day before being returned to their normal accommodation. Once the rat had reached the end location it was prevented (by a one-way door) from re-entering the body of the maze. Sufficient time was allowed in the end location to prevent any reward effects associated with food availability in their normal accommodation. On the eleventh day (i.e., after 11 runs through the maze) the second group were given access to food reward in the end location. A third, control, group was allowed to run the maze with no food reward throughout the duration of the experiment.

Tolman found that the performance of the second group on the twelfth daily run (the first after the introduction of reward) was as good as or better than that on the first group that had been rewarded on every run, who had shown a gradual

improvement in performance. Tolman's maze was constructed from 14 multiple T units, with doors between the units to prevent the rats retracing their steps in the maze. Tolman interpreted this as clear evidence that reward was not required for learning to take place. Tolman and Honzik's results are reproduced in figure 6-21. The measure of performance is the number of errors made by the experimental animal in traversing the maze.



Graphic from monolith\latent.cdr

Figure 6-21: Tolman and Honzik's Latent Learning Results

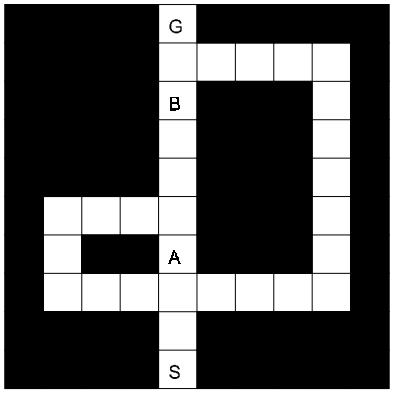
adapted from Bower and Hilgard (1981, p. 338)

6.6.1. Description of Procedure

A latent learning schedule is available as a fixed procedure in the SRS/E program. Figure 6-22 shows the experimental environment selected for this investigation. It is characterised by having three distinct paths of varying length from the defined start "S" to defined goal or finishing location "G". The maze arrangement used here differs from that of Tolman and Honzik.

In the procedure 100 "clone" animats are selected for each of the three groups (i.e., each of the three groups comprises 100 individuals with rseed = 1000, 1001 ... 1099). All 16 traversals of the maze by the first group are valenced. The first 11 traversals of the second group are unvalenced, but the twelfth and subsequent traversals are. All traversals by the control group are unvalenced. The essential

parameters are: Arep = 0.5, Adisp = 1.0, Lprob = 0.25, the other learning parameters are standard.



Graphic 5.25 from monolith\mazes.cdr

Figure 6-22: The SRS/E Latent Learning Environment

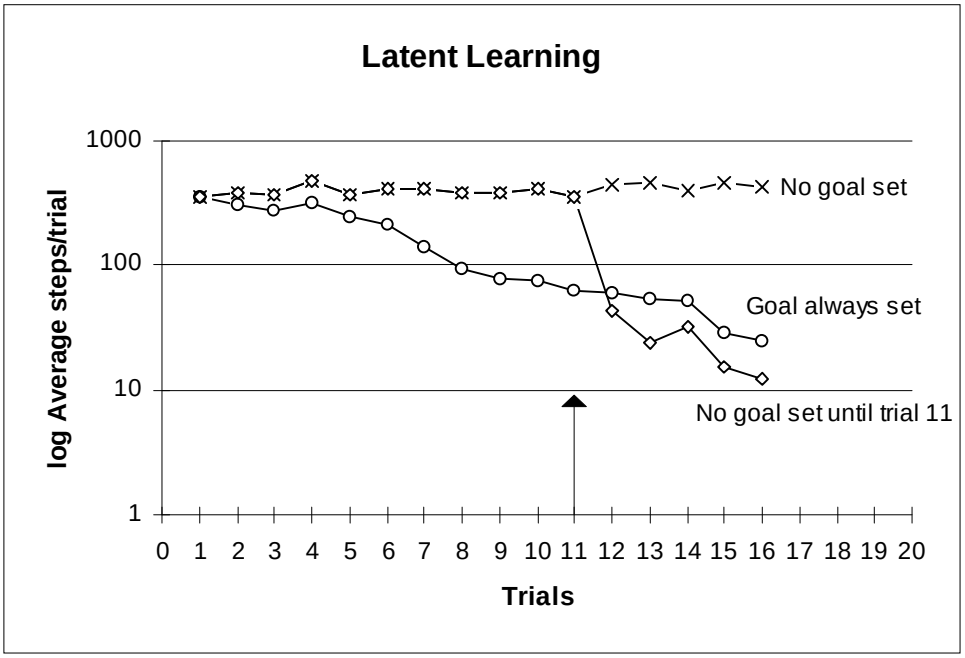
6.6.2. Results and Analysis of Experiment

Figure 6-23 shows the results of the experiment, indicating that the essential properties of the Tolman and Honzik experimental results are present. The first group show a gradual improvement in performance throughout the procedure. The second group show a dramatic improvement following the introduction of goal valencing. The third, control, group shows no significant change in performance. Note the different representation of performance, steps/trial rather than errors. A logarithmic representation of the performance axis has been used for cosmetic reasons. Neither of these factors should materially affect the interpretation of the results.

The gradual improvement seen in the control group of Tolman and Honzik's results is not replicated by SRS/E. This might be interpreted as evidence that some other form of reinforcement is available to the animal prior to the main reward (Bower and Hilgard, 1981, p. 339). Alternatively it might be noted that rats (and many other mammals) show a quite distinct *curiosity*[29], seeking out the novel and then ignoring it once it is no longer novel. The design of Tolman and Honzik's maze has many dead-ends, which once discovered can be safety ignored in

[29]As MacCorquodale and Meehl (1953, p. 204) put it "*No one who has observed rats during their early exposure to a maze could dismiss the exploratory disposition as of negligible strength*".

subsequent traversals of the maze - leading to a reduction in measured error rate. SRS/E differs in that it responds to novelty, but does not seek it out. An additional mechanism, such as *prioritized sweeping* of Moore and Atkeson (1993), might be adapted for use in SRS/E to demonstrate the gradual improvement findings in the control group.



monolith\results\latent\lat100.xls

Figure 6-23: Results of the SRS/E Latent Learning Experiment

6.6.3. Discussion

That SRS/E should demonstrate latent learning is hardly in doubt, nor a surprise. Reinforcement is generated internally, and is not dependent on external reward. Given the revival of interest in behaviourist and reinforcement learning methods for machine learning models it is nevertheless a timely reminder that these are well-trodden paths. Latent learning has been extensively researched. Thistlethwaite (1951) identifies and evaluates over 30 different latent learning experiments under a variety of different experimental conditions. MacCorquodale and Meehl (1953) placed considerable emphasis on the latent learning phenomenon, indeed stating that it provided the main motivation to add their contribution toward the formalisation of expectancy theory. MacCorquodale and Meehl note that not all experiments to demonstrate latent learning actually do so, in part, no doubt, due to variations in experimental design and procedure. Observation of the latent learning

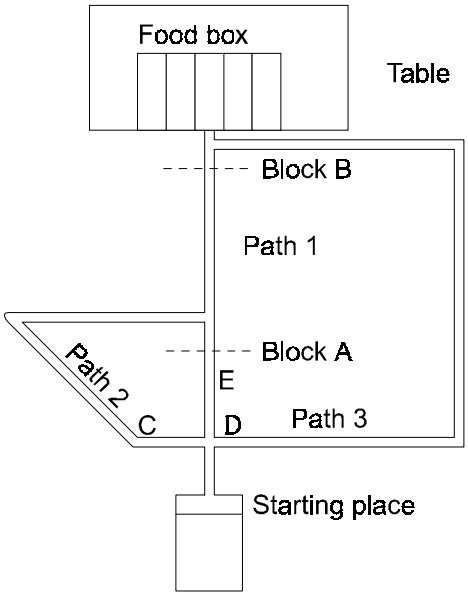
phenomenon places a considerable strain on behaviourist and reinforcement based theories, whereas the absence of the phenomenon has little impact on expectancy based models.

SRS/E's demonstration of the latent learning phenomena arises from one by now well explored propensity - to pursue a known route to a valenced goal in preference to exploring for a possible better alternative. With group one (always valenced) some, typically small, proportion of the individuals traverse the maze to the goal location by one of the longer paths during the first trial. Once they have that path, those individuals tend to continue to use it, as their behaviour is always valenced while in the maze. Gradual improvement in performance is a consequence of the choice of Lprob = 0.25, and is consistent with the learning rates previously shown in the baseline investigations of figure 6-2. Group two has adequate opportunity to explore the maze by random walk during the 11 unvalenced trials. Once the goal location becomes valenced individual animats have invariably encountered, and so use, the shortest route. Consequently, on average, the performance of group two exceeds that for group one, once the goal is valenced. The control group have no reason to treat the "goal" differently from any other location, and show no performance improvement.

6.7. Place Learning

Tolman also devised a *place learning* experiment, again using rats in an experimental maze to demonstrate what he referred to as "inferential expectation" or "insight" in these animals (Tolman and Honzik, 1930b). In this classic demonstration experimental rats were placed in a maze of the form shown in figure 6-24. With adequate experience of the maze rats show a clear preference for the shorter of the available routes, path 1. When path 1 was blocked the rats showed a distinct preference for path 2 and when path 2 was also blocked, then the rats would adopt the longer path 3. The key to the experiment is the placing of the block. Tolman argued that if the block was placed at point B a rat guided by blind habit would first try path 2, its choice at this decision point being directed by the response previously associated with the stimulus at that point. However, one capable of cognitive "inferential expectation" or "insight" would conclude that the

block also affected path 2 and would consequently employ path 3 directly. He found this to be the case.



Graphic 5.27 from monolith\tolmaze.cdr

Figure 6-24: Tolman and Honzik's "Insight" Maze

adapted from Bower and Hilgard (1981, p. 337)

6.7.1. Description of Procedure

These "insight" experiments are replicated with SRS/E using the experimental environment of figure 6-22. The procedure replicates the major functional features of Tolman and Honzik's "insight" maze. In the replication of this experiment naïve animats are allowed to explore the maze for 2,000 cycles by unvalenced random walk. This allows sufficient time for the animats to explore every path. Each animat is then given one valenced trial from "S" ("G" asserted as goal) with no path blocked to confirm that the animat will select the most direct route. In the next step the location at point "A" is blocked. The animat is returned to "S" and "G" is valenced. The number of steps required to traverse the environment to the goal is noted. The animat is returned to "S", "G" is valenced and the number of steps required to reach the goal location again noted. In the next step the block at location "A" is removed and a block added at location "B", the animat is returned to "S". The goal location "G" is valenced and the number of steps to traverse the modified environment noted. The animat is returned to "S" and the number of steps

to complete another valenced traversal to the goal location again noted. This experiment uses the standard learning parameters and Arep = 0.5, Adisp = 1.0, Lprob = 1.0.

6.7.2. Results and Analysis of Experiment

Figure 6-25 shows the performance in this experimental procedure by a single individual (rseed = 10). Sub-figure 6-25a confirms that each path has been fully explored, though by no means evenly. Sub-figure (b) confirms the animat takes the direct route when "G" is valenced. Sub-figure (c) shows the effect of the first valenced run after block "A" is set. After 10 failed attempts to traverse path 1, the animat proceeds along path 2, as Tolman would have predicted. Sub-figure (d) confirms the new path on the next valenced run. Sub-figure (e) shows the effect of the first valenced run after block "A" is cleared and block "B" set. As the animat is valenced it follows the known available route (via path two) until the unexpected block is encountered at "B". After a number of failed attempts to traverse the now blocked location "B" the animat backtracks down path one and round to the goal location via path three. The still longer path involving path two is ignored. Sub-figure (f) confirms the new route via path 3 on the next valenced run.

(A) 2000 steps random walk (seed = 10)

```
Cycle 2001: World is 10 by 10
**** **** **** ****  337 **** **** **** **** ****
**** **** **** ****   55    5    3   22   15 ****
**** **** **** ****   84 **** **** ****    6 ****
**** **** **** ****   73 **** **** ****    1 ****
**** **** **** ****   84 **** **** ****   19 ****
****   25    4   21  119 **** **** ****   18 ****
****    6 **** ****  141 **** **** ****    6 ****
****   92   51   77   70   67   64    7   33 ****
**** **** **** ****  221X**** **** **** **** ****
**** **** **** ****  198 **** **** **** **** ****
Mean = 64.133331, std dev = 75.047981
```

(B) Confirm Path 1

```
Cycle 2011: World is 10 by 10
**** **** **** ****    1X**** **** **** **** ****
**** **** **** ****    1    0    0    0    0 ****
**** **** **** ****    1 **** **** ****    0 ****
**** **** **** ****    1 **** **** ****    0 ****
**** **** **** ****    1 **** **** ****    0 ****
****    0    0    0    1 **** **** ****    0 ****
****    0 **** ****    1 **** **** ****    0 ****
****    0    0    0    1    0    0    0    0 ****
**** **** **** ****    1 **** **** **** **** ****
**** **** **** ****    1 **** **** **** **** ****
Mean = 0.333333, std dev = 0.483046
```

(C) Add Block "A"

```
Cycle 2036: World is 10 by 10
**** **** **** ****    1X**** **** **** **** ****
**** **** **** ****    1    0    0    0    0 ****
**** **** **** ****    1 **** **** ****    0 ****
**** **** **** ****    1 **** **** ****    0 ****
**** **** **** ****    1 **** **** ****    0 ****
****    1    1    1    1 **** **** ****    0 ****
****    1 **** **** **** **** **** ****    0 ****
****    1    1    1   10    0    0    0    0 ****
**** **** **** ****    1 **** **** **** **** ****
**** **** **** ****    1 **** **** **** **** ****
Mean = 0.862069, std dev = 1.800383
```

(D) Confirm Path 2

```
Cycle 2052: World is 10 by 10
**** **** **** ****    1X**** **** **** **** ****
**** **** **** ****    1    0    0    0    0 ****
**** **** **** ****    1 **** **** ****    0 ****
**** **** **** ****    1 **** **** ****    0 ****
**** **** **** ****    1 **** **** ****    0 ****
****    1    1    1    1 **** **** ****    0 ****
****    1 **** **** **** **** **** ****    0 ****
****    1    1    1    1    0    0    0    0 ****
**** **** **** ****    1 **** **** **** **** ****
**** **** **** ****    1 **** **** **** **** ****
Mean = 0.551724, std dev = 0.525226
```

(E) Remove Block "A", Add Block "B"

```
Cycle 2098: World is 10 by 10
**** **** **** ****    1X**** **** **** **** ****
**** **** **** ****    1    1    1    1    1 ****
**** **** **** **** **** **** **** ****    1 ****
**** **** **** ****   15 **** **** ****    1 ****
**** **** **** ****    2 **** **** ****    1 ****
****    1    1    1    2 **** **** ****    1 ****
****    1 **** ****    1 **** **** ****    1 ****
****    1    1    1    2    1    1    1    1 ****
**** **** **** ****    1 **** **** **** **** ****
**** **** **** ****    1 **** **** **** **** ****
Mean = 1.586207, std dev = 2.559633
```

(F) Confirm Path 3

```
Cycle 2116: World is 10 by 10
**** **** **** ****    1X**** **** **** **** ****
**** **** **** ****    1    1    1    1    1 ****
**** **** **** **** **** **** **** ****    1 ****
**** **** **** ****    0 **** **** ****    1 ****
**** **** **** ****    0 **** **** ****    1 ****
****    0    0    0    0 **** **** ****    1 ****
****    0 **** ****    0 **** **** ****    1 ****
****    0    0    0    1    1    1    1    1 ****
**** **** **** ****    1 **** **** **** **** ****
**** **** **** ****    1 **** **** **** **** ****
Mean = 0.620690, std dev = 0.491304
```

monolith\figures.ppt:slide 10

Figure 6-25: Results from "Insight" Experiment

As with the latent learning experiment the key to successful demonstration of the phenomenon under investigation is careful experimental layout and procedure. Where the latent learning procedure called for careful rationing of experience in the maze during the initial stages of the sequence, this procedure calls for adequate exploration. Without this the various routes may not be fully known to the animat, and consequently it will not select the preferred (by the experimenter in this case) routes. Other researchers subsequently found Tolman and Honzik's results repeatable, but prone to disruption, apparently due to elements in experimental design.

6.7.3. Discussion

SRS/E confirms Tolman's view of "insight". It seems unlikely that Tolman will have won much approbation from his peers by the use of the term, implying as it does, a level of intelligence well above that normally associated with the laboratory

rat. Perhaps paradoxically, and with the benefit of hindsight, we may see that this behaviour is fully explicable in terms of problem solving, at best a minor form of "insight". Nevertheless, the capabilities demonstrated by Tolman's rats and replicated by the SRS/E algorithm in this procedure still present considerable difficulties to the behaviourist and reactive agent schools of thought that promote reinforcement learning by explicit reward.

6.8. Chapter Summary

This chapter has described a series of experiments that investigate the properties of the SRS/E algorithm as an implementation of the Dynamic Expectancy Model. To facilitate direct comparison with previously published algorithms, Sutton's (1990) Dyna family of reinforcement learning programs, the experimental conditions employed for those previously published works have been replicated. In the baseline investigations of section 6.2 the performance of the SRS/E algorithm was directly compared to that of Sutton's *Dyna-PI* algorithm. SRS/E shows a marked performance gain over Sutton's algorithm. Under "ideal learning conditions" SRS/E was clearly able to master the maze traversal problem within a single trial (the Lprob = 1.0 curve of figure 6-2), whereas Dyna-PI is recorded as requiring over 80 trials (the "zero planning steps" curve of figure 6-1). It may be estimated that this represents approximately a forty-fold improvement in learning efficiency, in terms of the overall number of steps required to master the given task. The improved curves shown for Dyna-PI are achieved by added internal computation, the degraded curves for SRS/E are created by restricting the effectiveness of the learning process (Lprob < 1.0).

Sutton did not report on the performance of Dyna-PI in the noise disrupted environment he described. However, these investigations were performed with the SRS/E algorithm, and are reported in section 6.3. The results obtained are summarised in figures 6-4 and 6-5. The figures demonstrate that while the rate at which the task is learned is not markedly affected by the addition of this form of noise, the overall learned task performance is degraded by the presence of the noise. It was subsequently argued in section 6.3.2.2 that the Dynamic Policy Map is indeed correctly formed by the learning process. It is the task performance that is

disrupted by the presence of noise in the test trials. When this noise is removed, animat task performance is restored to near optimal levels.

The alternative and multiple goal experiments described in section 6.4 highlight a significant difference between the Dynamic Expectancy approach and that of conventional *Q*-learning algorithms. By recomputing the policy map on demand it becomes clear that any sign known to the system may be treated as a goal and selected on some arbitrary basis, not just those signs that were assigned as goals during the learning process. The SRS/E algorithm may therefore address situations where the animat is faced with goals that vary over time, and where several goals, of varying priority, must be tackled in an appropriate order.

The investigations of section 6.5 explored the response of the SRS/E algorithm to a variety of situations in which different paths from a starting point to a fixed goal point are presented to the animat. These tasks are essentially beyond the capabilities of conventional *Q*-learning algorithms of the form described by Watkins (1989). The performance of Sutton's *Dyna-Q+* algorithm, an adaptation of the *Q*-learning approach, was compared directly with the unmodified form of the SRS/E algorithm. Even though the mechanism by which new paths are discovered is radically different in the two algorithms, the apparent recorded performance was generally very similar. This is something of a surprise, as it might be thought that the inclusion of a continuously active exploratory component in the Dyna-Q+ algorithm would degrade its otherwise optimal levels of performance. Exploration is only invoked in SRS/E when an obstruction to the policy map path is encountered. The provision of an extinction mechanism in the SRS/E algorithm is a radical departure from the Dyna approach, and has some biological plausibility.

The demonstration of latent learning, described in section 6.6, highlights a substantive difference between the Dynamic Expectancy Model and previous reinforcement learning techniques. Learning is demonstrated to take place in the absence of external reward. This result, for which there is a substantial body of corroborating literature from animal learning experiments, would be wholly unexpected from a conventional reinforcement learning mechanism.

Similarly the place learning experiments, described in section 6.7, demonstrate the ability of the SRS/E algorithm to negotiate obstructions in its policy path in a manner that would be unpredicted from any algorithm employing a static policy map. Again, these results are consistent with findings from well-established animal learning experiments.