

**Schemes for Learning and Behaviour:  
A New Expectancy Model**

**Ph.D. Thesis**

**Christopher Mark Witkowski**

**February 1997**

**Department of Computer Science  
Queen Mary Westfield College**



**University of London**

## Abstract

This thesis presents a novel form of learning by reinforcement. Existing reinforcement learning algorithms rely on the provision of external reward signals to drive the learning algorithm. This new algorithm relies on reinforcing signals generated internally within the algorithm. The algorithm, SRS/E, described here generates expectancies ( $\mu$ -hypotheses), each of which gives rise to a specific prediction when the conditions relevant to the expectancy are encountered (the  $\mu$ -experiment). The algorithm subsequently tests these predictions against actual events and so generates reinforcement signals to corroborate or reject individual expectancies. This procedure allows for self-contained, completely unsupervised learning to an extent not possible with previous reinforcement procedures. The SRS/E algorithm is derived from a number of postulates that constitute a new Dynamic Expectancy Model developed in this thesis.

In contrast to the static policy map generated by existing  $Q$ -learning based reinforcement algorithms, which limit learning to one goal, the SRS/E algorithm generates a Dynamic Policy Map (DPM) from learned expectancies whenever a new goal is selected by the system. This new approach retains the advantages of reactivity to the environment inherent in existing reinforcement algorithms, while substantially increasing the system's flexibility in responding to varying circumstances and requirements. Also in contrast to previous reinforcement systems, goals may be selected arbitrarily and are not limited to those which were associated with reward during the learning steps. This new method allows multiple goals to be pursued either simultaneously or sequentially.

The single SRS/E implementation has been compared directly to the published results from of a family of reinforcement based algorithms, Dyna-PI, Dyna-Q and Dyna-Q+ (Sutton, 1990), themselves extensions to the groundbreaking  $Q$ -learning algorithm (Watkins, 1989). Under equivalent "ideal learning conditions" the SRS/E algorithm was found to outperform the equivalent Dyna reinforcement program to learn a simple maze task by a factor of some 40:1. The SRS/E learning algorithm was also found to be robust when tested under controlled "noise" conditions. SRS/E was also compared directly to Sutton's Dyna-Q+ algorithm on a range of alternative path and route blocking tasks and was found to offer a similar performance, but SRS/E employs a "biologically plausible" extinction mechanism, mirroring findings from animal behaviour research.

Finally SRS/E was tested with experimental designs for "latent learning" and "place learning", drawn directly from animal learning research. Both are regarded as presenting severe challenges to conventional reinforcement learning theories. SRS/E performs well on both tasks, and in a manner consistent with findings from animal experiments.

# Table of Contents

1. The Behaviour of Animals and Animats .....	9
1.1. Three Components of Natural Intelligence.....	10
1.2. Reactive Models of Intelligence.....	11
1.3. Action Selection Mechanisms.....	12
1.4. Arriving at a Definition of Learning.....	18
1.4.1. What is Not Learning.....	19
1.5. A Caveat.....	20
1.6. Thesis Outline .....	21
2. Theories of Learning.....	23
2.1. Classical Conditioning and Associationism .....	24
2.2. Reinforcement Learning .....	26
2.3. Computer Models of Reinforcement Learning .....	29
2.3.1. Markov Environments .....	31
2.4. <i>Q</i> -learning.....	32
2.4.1. <i>Q</i> -learning - Description of Process .....	33
2.4.2. Some Limitations to <i>Q</i> -learning Strategies .....	34
2.5. Classifier Systems.....	35
2.6. Artificial Neural Networks .....	38
2.7. Operant Conditioning.....	42
2.8. Cognitive Models of Learning, Tolman and Expectancy Theory .....	44
2.9. MacCorquodale and Meehl's Expectancy Postulates .....	46
2.10. Computational Models of Low-level Cognitive Theories .....	48
2.11. Becker's JCM Model .....	49
2.12. Mott's ALP Model.....	51
2.13. Drescher's Model.....	53
2.14. Other Related Work .....	55
3. A New Dynamic Expectancy Model.....	57
3.1. The Animat as Discovery Engine - The Thesis.....	58
3.2. The Expectancy Unit as Hypothesis.....	59
3.2.1. The Hypothesis Postulates .....	60
3.2.2. Initial Conditions for the $\mu$ -Hypothesis Set.....	63
3.2.3. Concluding Conditions for the $\mu$ -Hypothesis Set.....	63
3.2.4. Hypothesis Based Models of Learning .....	65
3.2.5. The Role of the Hypothesis in the Discovery Process.....	67
3.3. Tokens, Signs and Symbols.....	69
3.3.1. The Sign and Token Postulates.....	70
3.3.2. Initial Conditions for the Token and Sign Sets .....	71
3.3.3. Supporting Evidence for Signs and Tokens .....	71
3.4. Actions and Reification .....	73
3.4.1. The Action Postulates.....	73
3.4.2. Initial Conditions for Actions.....	74
3.4.3. Supporting Evidence for an Action Vocabulary.....	75
3.5. Goal Definitions .....	76
3.5.1. The Goal Postulates.....	76
3.5.2. Goals, Starting Conditions and Discussion .....	77
3.6. On Policies and Policy Maps .....	78

3.6.1. Policy Map Postulates.....	78
3.6.2. Evidence for Chaining.....	80
3.6.3. Evidence for Goal Suspension and Extinction .....	81
3.6.4. Comparison to <i>Q</i> -learning.....	83
3.7. Innate Behaviour Patterns .....	83
3.7.1. Balancing Innate and Learned Behaviour .....	85
3.8. Advances Introduced by the Dynamic Expectancy Model .....	86
4. The SRS/E Algorithm.....	89
4.1. Encoding the Ethogram: SRS/E List Structures.....	90
4.1.1. List Notation .....	91
4.1.2. Summary of Lists.....	93
4.2. Tokens and the Input Token List.....	94
4.2.1. Input Token List Values .....	95
4.3. Signs and the Sign List.....	96
4.3.1. Representing Signs .....	97
4.3.2. Other Sign List Values.....	99
4.4. Actions and the Response List.....	100
4.4.1. Response List Values.....	101
4.5. Innate Activity and the Behaviour List.....	102
4.5.1. Behaviour List Structure and Selection .....	102
4.5.2. Behaviour List Values.....	104
4.6. Goals and the Goal List.....	104
4.7. The Hypothesis List .....	105
4.7.1. Other Hypothesis List Values.....	107
4.8. Corroborating $\mu$ -Hypotheses, Predictions and the Prediction List.....	109
4.8.1. Prediction List Element Values .....	111
4.9. The Dynamic Policy Map (DPM) .....	112
4.9.1. Selecting actions from the DPM.....	115
4.9.2. Recomputing the DPM .....	116
4.9.3. The DPM, A Worked Example .....	117
4.9.4. Pursuing Alternative Goal Paths.....	121
4.9.5. Pursuing a Goal to Extinction .....	124
4.10. Creating New $\mu$ -Hypotheses .....	126
4.10.1. Maintaining the Hypothesis List.....	128
4.11. The SRS/E Execution Cycle.....	130
4.11.1. Summary of Execution Cycle Steps.....	131
4.12. The SRS/E Algorithm in Detail .....	134
4.12.1. Step 1: Processing Input Tokens and Signs .....	135
4.12.2. Step 2: Evaluating $\mu$ -Experiments on the Basis of Prior Prediction.....	136
4.12.3. Step 3: Selecting Innate Behaviours and Setting Goals.....	136
4.12.4. Step 4: Building the Dynamic Policy Map .....	138
4.12.5. Step 5: Selecting a Valenced Action.....	141
4.12.6. Step 6: Performing an Action.....	142
4.12.7. Step 7: Conducting $\mu$ -Experiments.....	142
4.12.8. Step 8: Hypothesis Creation and Management .....	143
4.13. Implementation .....	146
4.14. SRS/E - A Computer Based Expectancy Model.....	146
5. Experimental Design and Approach .....	148

5.1. Experimental Design .....	149
5.2. The User Interface .....	151
5.2.1. Controlling Execution Cycles.....	152
5.2.2. Displaying and Recording List Information .....	153
5.2.3. Managing Goals.....	153
5.2.4. Managing the Animat and Environment .....	153
5.2.5. Accessing Utilities .....	154
5.3. The System Execution Trace Log.....	155
5.3.1. Processing Log Results.....	155
5.4. Important Schedule Variables.....	156
5.4.1. Action Repetition Rate (Arep) .....	156
5.4.2. Action Dispersion Probability (Adisp).....	157
5.4.3. Learning Probability Rate (Lprob) .....	158
5.5. Fixed Schedule Experiments.....	158
6. Investigations and Experimental Results.....	160
6.1. The Individual Experiments.....	162
6.2. Baseline Investigations .....	163
6.2.1. Description of Procedure .....	164
6.2.2. Results and Analysis of Baseline Experiment.....	165
6.2.3. Discussion .....	169
6.3. The Effects of Noise.....	171
6.3.1. Description of Procedure .....	171
6.3.2. Results and Analysis of Experiment .....	171
6.3.2.1. Tuning Parameters for Static Environments.....	172
6.3.2.2. The Effects of Noise: Learning or Behaviour? .....	175
6.3.3. Discussion .....	177
6.4. Alternative and Multiple Goals .....	178
6.4.1. Description of Procedure .....	178
6.4.2. Results and Analysis of Experiment .....	179
6.4.3. Discussion .....	183
6.5. Multiple-Path, Blocking, Shortcut and Extinction Investigations.....	184
6.5.1. Investigation One (Multiple-Path), Procedure .....	185
6.5.2. Investigation One, Results and Analysis .....	186
6.5.3. Investigation One, Discussion .....	188
6.5.4. Investigation Two (Goal Extinction), Procedure .....	188
6.5.5. Investigation Two, Analysis of Results.....	189
6.5.6. Investigation Two, Discussion .....	192
6.5.7. Investigation Three (Path Blocking), Procedure .....	193
6.5.8. Investigation Three, Results and Analysis.....	194
6.5.9. Investigation Three, Discussion .....	196
6.6. Latent Learning.....	197
6.6.1. Description of Procedure .....	198
6.6.2. Results and Analysis of Experiment .....	199
6.6.3. Discussion .....	200
6.7. Place Learning .....	201
6.7.1. Description of Procedure .....	202
6.7.2. Results and Analysis of Experiment .....	203
6.7.3. Discussion .....	204
6.8. Chapter Summary .....	205

7. Extensions to SRS/E and Further Work .....	208
7.1. An Association List .....	208
7.2. Seeking Multiple Goals Simultaneously .....	209
7.3. An Explicit Template List.....	211
7.4. Directing Learning Effort .....	212
7.5. Aversion .....	214
8. Discussion and Conclusions .....	216
8.1. Reactive or Cognitive? .....	216
8.2. Expectancy Model as “Missing Link” in Learning Theory .....	216
8.2.1. Types of Reinforcer .....	217
8.3. Relationship to Policy Maps and Universal Plans .....	218
8.4. One-Shot Learning Phenomena .....	221
8.5. Expectancy Theory and XBL - a Proposal.....	222
9. Appendix One.....	224
10. References .....	229
11. SUBJECT and AUTHOR INDEX .....	251

## List of Figures and Tables

1-1: Tinbergen’s Principle of Hierarchical Organisation.....	14
1-2: Brooks’ Subsumption Architecture.....	15
1-3: Maes’ Action Selection Architecture .....	17
2-1: A Classifier System.....	36
2-2: A Simple Neurone Model .....	39
2-3: A Multilayer Neural Network Model .....	41
2-4: A JCM Schema .....	50
2-5: A Schema in Drescher’s Cognitive Model.....	53
2-6: A Composite Action.....	54
2-7: The Marginal Attribution Process .....	55
3-1: Extinction Curves Under Various Schedules.....	82
Table 4-1: SRS/E Internal Data Structures.....	92
4-1: Log Printout of a Valenced Path.....	115
4-2: Model DPM Generated from Sample Hypothesis List.....	118
4-3: Various Outcomes for Model DPM .....	119
4-4: Model Graph Recomputed for Goal ‘S8’ .....	121
4-5: A Sample Dynamic Policy Map .....	122
Table 4-2: Paths Through Figure 4-5 Graph.....	123
4-6: Summary of Steps in the SRS/E Execution Cycle .....	132
4-7: The SRS/E Algorithm.....	134
4-8: Step One, Token and Sign Processing.....	135
4-9: Step Two, Evaluation of $\mu$ -Experiments .....	136
4-10: Step Three: Select Innate Actions and Set Goals.....	137
4-11: Step Four, Construct Dynamic Policy Map .....	140
4-12: Step Five, Select Valenced Action .....	141
4-13: Step Six, Perform Action.....	142
4-14: Step Seven: Conduct $\mu$ -Experiments.....	142
4-15: Step Eight, Hypothesis Creation .....	144
4-16: Step Eight, Hypothesis Management - Specialisation .....	145
4-17: Step Eight, Hypothesis Management - Forgetting .....	146
5-1: Sutton’s DynaWorld/Standard Environment .....	149
5-2: The SRS/E Experimenter Command Options.....	152
5-3: Effect of $A_{\text{rep}}$ on Random-Walk Path Length.....	157
6-1: Results from Sutton’s Dyna-PI Experiments .....	164
6-2: Baseline Learning Curves ( $L_{\text{prob}} = 1.0, 0.25, 0.1$ and $0.025$ ).....	166
6-3: Contribution of Individual Animats to Learning Curve.....	168
6-4: Baseline Learning with Noise ( $A_{\text{disp}} = 0.5, L_{\text{prob}} = 1.0, 0.25, 0.1$ and $0.025$ ).....	172
6-5: Baseline with Noise ( $A_{\text{disp}} = 0.5, \gamma^1 = 1.0$ ).....	174
6-6: a) Path with $A_{\text{disp}} = 0.5$ (trial 101), b) $A_{\text{disp}} = 1.0$ (trial 102).....	175
6-7: Policy Map at Conclusion of Trial 101 .....	176
6-8: Planned Valenced Path (trial 101).....	177
6-9: Simultaneous Goal Locations .....	179
Table 6-1: Results for Investigation One of Dual Goal Experiment.....	180

6-10: Animat Random and Valenced Paths (investigation 1, rseed = 80) .....	181
Table 6-2: Results for Investigation Two of Dual Goal Experiment.....	182
Table 6-3: Results for Investigation Three, Simultaneous Goals .....	183
6-11: Sample Simultaneous Goal Paths .....	183
6-12: Changing World Environments .....	185
6-13: Multiple Path Investigation, Individual Performance .....	186
6-14: Estimated Cost Profile (Path and H14) .....	187
6-15: Goal Extinction .....	190
6-16: Goal Extinction, Comparison of Cost Estimate to VBP.....	191
6-17: Goal Extinction (two path), Cost Estimate and VBP.....	192
6-18: Average Performance of Dyna Systems on a Blocking Task.....	194
6-19: Investigation Three, Individual ‘Cumulative Reward’ Curves.....	195
6-20: Investigation Three, Average ‘Cumulative Reward’ Curves .....	196
6-21: Tolman and Honzik’s Latent Learning Results.....	198
6-22: The SRS/E Latent Learning Environment .....	199
6-23: Results of the SRS/E Latent Learning Experiment .....	200
6-24: Tolman and Honzik’s ‘Insight’ Maze .....	202
6-25: Results from ‘Insight’ Experiment .....	204
7-1: Sign-Sign Associations (Secondary Cathexis) .....	209
7-2: Enhanced Goal Acquisition.....	210
7-3: The Effect of Valence Level Pre-Bias .....	213