Chapter 3

3. A New Dynamic Expectancy Model

This chapter seeks to define and develop a new *Dynamic Expectancy Model*. This Dynamic Expectancy Model extends MacCorquodale and Meehl's original expectancy theory formulation to provide a workable and so testable implementation. It may be seen as part of the current trend to identifying existing "thought experiments" from the literature, reconstructing them as computer simulations and so re-evaluating and reviewing their premises and predictions by experiment and analysis in a manner that was previously impossible. The Dynamic Expectancy Model builds on the intermediate level cognitive models described by Becker (1973), Mott (1981) and Drescher (1991). It also draws on mechanisms and processes from a range of other sources, notably the accumulated work on innate behaviours and capabilities (Tinbergen, 1951; Brooks, 1986; and Maes, 1991, among others) and the notion of a *policy map* drawn from reinforcement learning methods (Sutton, 1990; Watkins, 1989).

The Dynamic Expectancy Model eschews mechanisms exclusively detected in human infant or adult subjects, but serves rather to address issues arising from work relating to the understanding and modelling of animal behaviour. In particular this new model identifies and addresses some of the limitations and shortcomings of behaviourist theories relating to learning and behaviour in lower animals, which were considered in previous chapters. The new model focuses on the idea that all animals (of whatever level of complexity) are essentially autonomous individuals, which may behave, learn and reason within the capabilities ultimately determined by their innate definition, the *ethogram*. This individuality does not imply that those individuals exist independently of other members of the same or other species. Many are dependent on parental care, naturally exist and co-operate in packs or communities composed of distinct individuals, exist in symbiotic or antagonistic relationships, or must attract a mate to reproduce.

The intermediate-level cognitive models of Becker, Mott and Drescher seek to emulate the developmental process of the human infant. Each was influenced to varying extents by the work of the Swiss child developmental psychologist *Jean Piaget* (1896-1980). Drescher (1991, Ch. 2) provides a description of the first six stages of infant development according to Piaget's observations. One fundamental problem with this approach is the rapidity with which normal human infant development proceeds. These intermediate-level cognitive models lack the power to account for the considerable increases in the child's performance and ability. Moreover, there is still little agreement as to whether some, most, or all of this observable improvement is primarily due to a learning or to a maturation process in which innate abilities are activated in an essentially constant order. These models may therefore be taken as simplifications of other cognitive-organisational theories of learning (Bower and Hilgard, 1981, Ch. 13) which are obliged to postulate a wide range of mechanisms to account for the diversity of human adult abilities. Tolman and expectancy theory takes a constructivist view, adopting mechanisms required to model and explain behaviour and ability of non-human animals, though he later attempted to expand the model to encompass many aspects of human behaviour.

3.1. The Animat as Discovery Engine - The Thesis

In the Dynamic Expectancy Model animats may be viewed as machines for devising hypotheses, conducting experiments and subsequently using the knowledge they have gained to perform useful behaviours. In this learning model the animat implements a low level version of a "scientific discovery process." A critical feature is the creation and verification of self-testing experiments, derived from simple hypotheses created directly from observations in the environment. Each hypothesis describes and encapsulates a simple experiment. Each experiment takes the form of an expectancy or prediction that is either fulfilled, so corroborating the effectiveness of the hypothesis, or is not fulfilled. From time to time goals, activities required of the animat, will arise. By constructing a graph like structure from the hypotheses it has discovered during its lifespan and then determining an intersection of this graph with current circumstances, the animat may determine appropriate actions to satisfy those goals. Part of the innate structure of the animat provides the rules by which this discovery process

proceeds. Part imbues the animat with sufficient behaviour to set goals and to initiate and continue all these activities until learned behaviour may take over from the innate. Above all the animat must survive long enough to create hypotheses and conduct experiments.

Where Popper (1959, and see section 3.2.5 later in this chapter) describes a *Hypothetico-Deductive* approach, the Dynamic Expectancy Model adopts a *Hypothetico-Corroborative* stance. No mechanism for the construction of more complex models is incorporated into the Dynamic Expectancy Model. In order to distinguish hypotheses in the Dynamic Expectancy Model from those proposed by Popper, they will be referred to as *μ-hypotheses* ("micro-hypotheses"), similarly experiments as *μ-experiments* ("micro-experiments"). The construction and verification of low-level observation based μ-hypotheses would appear a useful pre-cursor to the independent development of any systematic theoretical model, whose structure is not wholly or primarily dependent on an *originator*[12].

3.2. The Expectancy Unit as Hypothesis

In the Dynamic Expectancy Model the expectancy, and so the basic unit of learning, takes the form of the predictive μ-hypothesis. This has critical implications. First and foremost of these implications is that each expectancy unit now contains the means to perform a self-contained test and so confirm or deny its own validity. In turn this implies the learning process is no longer dependant on external or reward signals to guide the process. Behaviour to seek goals is made independent of learning activity required to accumulate the knowledge, which may in turn be applied in performing goal directed behaviour. This section describes and discusses a number of "postulates" that define the operation of the expectancy unit as predictive hypothesis.

[12] Originator, the individual or process responsible for the creation of the animat and its ethogram.

3.2.1. The Hypothesis Postulates

Definition H0: The **μ-hypothesis**. Each μ-hypothesis records an assumed transition between two detectable sensory patterns (signs "s1" and "s2", q.v.) indicated or caused by an action ("r1") available to the animat system.

Postulate H1: **Prediction**. Prediction forms the basis of self-testability. Each μ-hypothesis encapsulates an expectation that predicts the occurrence (or appearance) of the consequent sign ("s2") at a specific time following the appearance (or occurrence) of the context sign ("s1") and the action ("r1").

Postulate H2: **μ-Experimentation**. μ-Experimentation is the mechanism by which predictive self-testability is achieved. Every μ-hypothesis is tested at every opportunity. A separate prediction relating to the consequent sign "s2" is created each and every instance where the context sign "s1" and response "r1" occur in the relationship defined in that μ-hypothesis. Each such prediction is termed a μ-experiment. The conduct of μ-experiments is insensitive as to why the triggering conditions "s1" and "r1" arose.

Postulate H3: **Corroboration**. Corroboration is one method by which the predictive ability of a μ-hypothesis is recorded. The quality of a μ-hypothesis is determined solely by its ability to accurately predict its consequent sign. The corroboration measure is defined as the ratio of the total number of predictions made by the μ-hypothesis to the number of correct predictions made, as verified *post-priori*. Any μ-hypothesis that has always given rise to a verified prediction will have a corroboration measure of 1.0. Any other μ-hypothesis will have a confidence or "corroboration" measure (Ch) of zero or greater, but less than one. Ch therefore reflects the probability of a valid prediction, thus:

$$Ch = p(s2 \mid^{t} s1+r1) \qquad \text{(eqn. 3-1)}$$

The use of the "*t*" symbol acts as a reminder of the temporal relationship that exists between the expectandum "s2" and the context. As this expression gives no indication of sample size, the corroboration measure is not in itself an indication of the usefulness, rarity or reliability of the prediction.

Postulate H4: **Reinforcement**. Reinforcement is a second method by which the predictive ability of a μ-hypothesis is recorded. In this context "reinforcement" substitutes for MacCorquodale and Meehl's use of the term *mnemonization*. In a measure related to corroboration, each successful verified prediction reinforces confidence in a μ-hypothesis. Conversely every unsuccessful prediction extinguishes confidence in that μ-hypothesis. The effect of each verification is discounted as further predictions are made. The *reinforcement measure* (Rh) is changed by the quantity:

$$\Delta Rh^{p+1} = \alpha(1 - Rh^{p}) \qquad \text{(eqn. 3-2)}$$

following each instance of a successful prediction (^{p}), and

$$\Delta Rh^{p+1} = - \beta(Rh^{p}) \qquad \text{(eqn. 3-3)}$$

following each unsuccessful prediction. Under constant conditions these relationships give rise to the widely observed "negatively accelerating" form of the learning curve. The two proper fractions the *reinforcement rate* (α) and the *extinction rate* (β) respectively define a "learning rate" for successful and unsuccessful prediction situations. They control the rate at which the influence of past predictions will be discounted. These parameters shall be normalised such that the Rh value of a μ-hypothesis that makes persistently successful predictions tends to 1.0, the Rh value of a μ-hypothesis that persistently makes unsuccessful predictions tends to 0.0. The positive reinforcement rate need not be equal to the negative extinction rate.

Mnemonization for expectancies in the MacCorquodale and Meehl postulates are fundamentally based on the notion of temporal adjacency and contiguity. This was inherited from decades of experimental observation that has repeatedly noted that learning phenomena are invariably stronger for events that are closely related in the temporal domain. This is entirely consistent with the provisions of the Dynamic Expectancy Model. Temporally adjacent predictions are tested first. The time-scale being extended only in circumstances where unsatisfactory predictive performance is determined over the shorter period.

Postulate H5: **Creation**. Creation is the method by which the animat extends the set of μ-hypotheses. μ-Hypotheses exist to predict future occurrences of signs; it is therefore reasonable to suppose that new μ-hypotheses might be created under two specific circumstances. First, every sign shall have at least one μ-hypothesis capable of predicting it. Novel signs (ones not previously recognised by the system) shall trigger a rule creation process, postulate H5-1, *novel event*. The consequence ("s2") for this new μ-hypothesis will be the novel sign. The context and action drawn from the set of recent signs and actions recorded by the system. By a process of *timebase shifting* the current, novel, sign will be shifted to be a future prediction, with a corresponding shift in the relative time relationship to the other components selected for the new μ-hypothesis.

In the second creation circumstance, known signs are detected without a corresponding prediction, postulate H5-2, *unexpected event*. A new μ-hypothesis may be created, using the same mechanism as for novel signs to cover the unexpected event. Shen (1994) and Riolo (1991) both describe broadly similar strategies for "rule" creation triggered by "surprise" events. Kamin (1969) has investigated the role of predictability and surprise in various classical conditioning procedures using rats.

Postulate H6: **Differentiation**. Differentiation is the mechanism by which the animat may refine its existing set of μ-hypotheses. Differentiation adds extra conditions to the context of an existing μ-hypothesis, reducing the range of circumstances under which that μ-hypothesis will be applicable. Differentiation may be appropriate to enhance μ-hypotheses that have stabilised, or stagnated, at some intermediate corroborative measure value. μ-Hypotheses should not be subject to differentiation until they have reached an appropriate level of testing (their "maturity"). Maturity is a measure of the degree of corroboration of a μ-hypothesis. It is otherwise independent of the age of a μ-hypothesis. It is expected that the differentiation process will create new, separate μ-hypotheses that are derived from the existing ones. Both old and new μ-hypotheses are retained and may then "compete" to determine which offers the best predictive ability.

Postulate H7: **Forgetting**. Forgetting is the mechanism by which the animat may discard μ-hypotheses found ineffective from the set of μ-hypotheses held. A μ-hypothesis might be deleted when it can be determined that it makes no significant contribution to the abilities of the animat. This point can be difficult to ascertain. Evidence from animal learning studies indicates that learned behaviours may be retained even after considerable periods of extinction. Experimental evidence from the implementation of the model described later will point to the value of not prematurely deleting μ-hypothesis, even though their corroborative measures fall to very low levels. Where a sign is predicted by many μ-hypotheses there may be good cause to remove the least effective. It is presumed that the last remaining μ-hypothesis relating to a specific consequent sign will not be removed; on the basis that some predictive ability, however poor, is better than none at all. Even if it was to be removed, a new μ-hypothesis would be created (by H5-2, unexpected event) on the first re-appearance of the consequent sign of the deleted μ-hypothesis. As no record is retained of the forgotten μ-hypothesis, any new μ-hypothesis created may be the same as one previously removed.

3.2.2. Initial Conditions for the μ-Hypothesis Set

The ethogram may be programmed to contain pre-determined μ-hypotheses, which will be used, corroborated, differentiated and forgotten as any other μ-hypothesis available to the animat. Equally the set of μ-hypotheses available to the animat may be empty at the time of *parturition*[13], the set being populated and maintained by actions defined by the various postulates described.

3.2.3. Concluding Conditions for the μ-Hypothesis Set

The animat is assumed to have a limited lifespan, but only by analogy with natural animals; there is no explicitly defined concluding or *terminating condition* defined in the Dynamic Expectancy Model. Learning by μ-hypothesis creation may slow and finally cease in the event that no new signs are encountered by the system, and when the existing signs are adequate to predict every appearance of each sign. These conditions may be encountered in the special environment defined by the

[13] Parturition, the moment the animat becomes a free-standing individual, dependent on the definition contained within the ethogram; analogous, perhaps, to the birth of an animal.

finite deterministic Markov state space environment (*FDMSSE*). Under these specific conditions, once every state has been visited at least once, then there will be no further μ-hypothesis creation on the basis of novelty (H5-1). Once every transition has been attempted in each state no new rules will be created on the basis of unpredicted appearance (H5-2). At this point there is a μ-hypothesis to accurately predict the next state, so that the conditions required to invoke μ-hypothesis differentiation (H6) and forgetting (H7) do not arise. Corroboration (H3/H4) does not cease under these conditions, neither does the option to recommence μ-hypothesis creation, differentiation or forgetting should the underlying structure of the environment change for any reason. It has been assumed that the animat has, inherent in its ethogram, some strategy that will eventually allow it to visit all states by all transition options. This may be by selecting actions at random.

A similar argument may be advanced in the case of the finite stochastic Markov state space environment (*FSMSSE*). As in the FDMSSE situation, learning by creation (H5-1) will cease once each state has been visited. Once each transition has been made, including all those derived from the additional probabilistic nature of the environment, creation by unpredicted event (H5-2) will cease. After an extended period of exploration in the environment the corroborative measure (H3) of each μ-hypothesis will tend to the true probability of the associated transition, although this will only ever be an estimate of the true probability. As before, should the structure of the state space change (new states or new transitions) new μ-hypotheses will be created to accommodate those changes.

Should the relative distribution of transition probabilities change, both the corroborative (H3) and reinforcement (H4) measures will change to reflect this as further exploration takes place. The corroborative measure reflects the overall "lifespan" situation. Under these circumstances the reinforcement measure has the potential to provide a better working estimate. Due to the probabilistic nature of the transitions none of the μ-hypotheses will achieve full corroboration. When the initial set of μ-hypotheses reaches the required level of maturity the differentiation process (H6) will become activated. New μ-hypotheses formed are subsequently tested in competition with their prototypes. Under the FSMSSE model conditions new context signs will be created by concatenation of additional states drawn from

recorded past states (only one state is indicated at the current time). Given that the definition of the FSMSSE restricts the information bearing content for the choice to the current state, it may be taken that all such μ-hypotheses created by differentiation will, in the limit, be less effective than their parent prototypes. It is therefore an unfortunate consequence of the basic assumptions of the FSMSSE that differentiation will continue throughout the animat's lifecycle, without materially improving its behavioural performance. On the other hand its effect will not be catastrophic, the majority of the behaviour being mediated by the better corroborated initial set of μ-hypotheses.

Note that neither in the postulates, nor in either of these discussion cases (FDMSSE and FSMSSE) has any reference been made to the provision of an external source of reinforcement.

In general, the Markov state space environment may be considered a poor model of the natural environment. The fundamental assumption that the information required to select the best action to take is, or can be, described by the current sensory pattern remains, at best, contentious. Equally the idea that some combination of sensations will completely and uniquely describe a "state" that is constant over time and so may be returned to on numerous occasions fails to reflect our notion or experience of the natural world. Nevertheless, the FDMSSE and FSMSSE environments represent a well defined and extensively studied formalisation. They represent a convenient, repeatable and controlled test environment in which to conduct experiments to determine the properties and performance of a learning system. As these environments have been utilised by other authors, the Markov description represents a point of comparison between alternative theories of learning. Later sections in this work will return to the utility of the Markov environment as a test environment, and to comparisons with other research that has used these environments.

3.2.4. Hypothesis Based Models of Learning

An early suggestion that rats exploring maze test environments use a form of hypothesis was proposed by Krechevsky (1933). The term was later adopted briefly by Tolman (1938) as a description of his basic expectancy unit, although in

his later writings the term “field-expectancy” is preferred. Restle (1962) provides a mathematical formalisation in which “hypotheses” (assumed or untested patterns of responses to cue stimuli) are sampled from a fixed size population by different means. In Restle’s model, hypotheses were either always correct (“C”), always wrong (“W”), or inconclusive (“I”), sometimes wrong, sometimes correct. Restle further proposed three selection strategies. Strategy (1) in which one hypothesis was selected and tested, then another, and so on (the *single-hypothesis assumption*). In strategy (2) all available hypotheses are selected for testing. In strategy (3) samples from the total population of available strategies are selected for testing (the *sub-set sampling assumption*). Restle was able to demonstrate that (under defined conditions) these three strategies are essentially equivalent - the “indifference to sample size” theorem.

Levine (1970) conducted a series of experiments with human subjects, designed to identify which strategy was used by the subjects. Subjects were asked to sort cards according to four easily discriminated elements (size, form, brightness and position). On some trials the subjects were given an indication, “right” or “wrong”, about their choice so that they may form one or more “hypotheses” about their selection choice (which may guide their future decisions). Interspersed with these indicated trials the subjects made unguided choices. Such *blind-trials* allow the experimenter to infer the hypotheses in use by the subject. These studies concluded that subjects repeated a hypothesis indicated as correct, and discarded a hypothesis indicated as incorrect. More significantly, many of the subjects appeared to be sampling several hypotheses at each stage, the sub-set sampling assumption, as indicated by the number of trials prior to perfect performance. In a related set of experiments the latency time for the choice was measured over successive trials. These experiments demonstrated a fall in decision time as possible, but ineffective, hypotheses were discarded. Decision latency time remained constant following the “solution trial”. More recent studies (Klahr, 1994) indicate that the hypothesis generation strategy used by human subjects is dependent on age and educational level. These results may call into question the appropriateness of applying data derived from human subjects directly to autonomous learning in animals or animats.

The emphasis of Kruchevsky's work was that rats explored their environments in a methodical, rather than random, trial-and-error, way. The basic assumption driving both Restle's and Levine's research was that hypotheses are selected and retained or rejected from a finite, known, set. In Levine's procedure subjects were apprised of the set size before the trials began. The Dynamic Expectancy Model makes no assumption about pre-existing sets of hypotheses. Hypotheses are generated and tested as the opportunity arises. In turn this gives rise to other possible μ-hypothesis creation (postulate H5) strategies. Implicit in the description so far is the idea that the animat initially creates a single, minimally simple hypothesis for each situation, tests that hypothesis for some while, and subsequently may need to refine or replace it. An alternative strategy might be to create a group of μ-hypotheses, utilising both the spatial and temporal aspects of the trace, and subsequently aggressively reject or delete all those from this sub-set that are not corroborated on subsequent trials, an "over-sampling" assumption. Under this assumption it may be appropriate that learned μ-hypotheses do not affect the behavioural repertoire until this initial selection phase is complete, leading to a flat section just prior to the main learning curve[14].

3.2.5. The Role of the Hypothesis in the Discovery Process

This thesis presents animal learning as a process of discovery. As part of the arguments leading to his development of the central thesis in his classic and seminal work into the nature of the scientific process, his "*Logic of Scientific Discovery*", the eminent Austrian born philosopher Sir Karl Popper (1902-1994) identified many essential properties of the hypothesis and its role in a self-sustaining discovery process encapsulated in a set of "methodological rules" (Popper, 1959). In this view of the discovery process "scientific truth" is determined by the creation of hypotheses, which are tested from the phenomena they predict. In turn experiments are devised to determine the validity of the prediction. This is a form of *modus tolens*[15], where theories from which hypotheses were properly derived are discarded when the hypotheses are falsified by experiment. While Popper

[14]Kleitman & Crisler (1927) present data showing a similar effect under classical conditioning conditions.

[15]If t, some theory, implies p, some conclusion (say a logically derived hypothesis), then the falsifying inference "$((t \rightarrow p).\neg p) \rightarrow \neg t$" requires us to reject t if we find p false.

decisively rejects inductive logic ("theory from examples"), he provides scant clue in these early writings as to how he considers theories themselves are to be formulated. Later authors active in the field of the philosophy of science have extended this model, and provided alternative views, of the scientific discovery process. Berkson and Wettersten (1984) have attempted to apply the principles of Popper's Logic of Discovery to the psychology of learning.

The "Logic of Scientific Discovery" (LSD) contains many insightful observations about the nature of the discovery process. A number of these observations, pertinent to expectancy theory and particularly relating to the nature of the hypothesis and experiments are considered now. Hypotheses that have more general applicability, those giving rise to a smaller range of derived "statements" and so have a higher "empirical content", have decreasing opportunity to escape *falsification* (LSD, s31). It is therefore incumbent on the discovery process to propose the simplest theories and hypotheses that are testable and so falsifiable, though simplicity itself is not a substitute for falsifiability. Hypotheses that are not testable ("undecidable" or "meta-physical") or those which are trivially true[16] ("tautologous") are to be discarded. Selection of the fittest systems of hypotheses should be as a result of the "*fiercest struggle for survival*" (LSD, s.6). Even if inadequate such systems of hypotheses should persist until falsified or replaced by one better able to be tested and found more fit.

Experiments are derived from, and test, hypotheses. Experiments must therefore encapsulate a complete description of the conditions under which the phenomena under test will be reproducible. Any conditions not included in the experimental procedure being considered irrelevant. In Popper's view a hypothesis may at best be corroborated, or otherwise falsified, and consequently the hypothesis and therefore the theory from which it was derived should be refined or refuted. In practice Popper recognises that there may be valid exceptions to the strict application of this approach, such as when the hypothesis fails due to incomplete specification, or where verifying observations have reached the limits of available experimental technique. In Popper's model of the scientific method hypotheses are

[16]It has subsequently become apparent that practical logic based systems which ignore the trivially true or apparently commonplace are prone to particularly gross omissions of reasoning (the "common-sense" component).

deduced from theories (the *Hypothetico-Deductive* approach). In the Dynamic Expectancy Model hypotheses are generated directly from observations and tested (the *Hypothetico-Corroborative* approach). In both schemes testing of hypotheses is a continuous process, the "*scientific game*" one without end. We may decide to suspend testing a hypothesis temporarily, but "*he who decides ... that scientific statements do not call for any further test, and that can be regarded as finally verified, retires from the game*" (LSD, s10).

Experiments are repeated so that we may "*convince ourselves that we are not dealing with a mere isolated coincidence*" (LSD, s.8). Popper refers to such coincidences as *occult occurrences*, repeated testing validates or rejects the phenomenon. A similar effect has been noted by experimental psychologists in animals, a behaviour based on a single rewarding circumstance, which persists even though the outcome is not repeated. This effect is usually referred to as *superstitious learning*, characterised as the elicitation of ritualistic or stereotyped behaviour under non-contingent "reward" schedules. Skinner (1948) describes an experimental schedule demonstrating the phenomenon in pigeons. Blackman (1974, Ch. 2) reviews "superstitious" behaviours in an operant conditioning context. This effect is apparently distinct from superstitious behaviour in humans, based on mystic or other beliefs (Jahoda, 1969).

3.3. Tokens, Signs and Symbols

Signs are specifically a combination of one or more elementary sensory units. They recognise a condition that may itself be composed of more than one sensory mode. In the Dynamic Expectancy Model these individual elements are referred to as *tokens*. Tokens perform the initial conversion of data from external transducers or sensors into symbolic form. Sensors abound in nature and it is not intended to further review the scope or extent of animal senses here. Similarly there have been significant advances in artificial transducers that may be incorporated into robotic devices. In the present model tokens will be represented as two-state symbols, indicating the presence or absence of the condition detected. This is a limitation that may need to be addressed in the future. The values of past tokens are recorded in an *activation trace*, specifically to allow *temporal discrimination*. By referring

to elements in the activation trace behaviours may be related to past events, as well as those which are current.

3.3.1. The Sign and Token Postulates

Definition T0: **Token**. A token is a symbol relating to a basic unit of sensory input. A token indicates the instantaneous output from a detector. In the present model a token is either active or inactive, reflecting one of two possible detector states. Tokens are time tagged. They may represent the state of the detector at the current time or provide a record of the state of each detector at given times from the recent past (the "activation trace"). Older token records are discarded. Tokens may be attached to transducers to detect physical aspects relating to the animat and its environment. Tokens may also detect information processing activities within the animat.

Definition S0: **Sign**. A sign encapsulates a combination of conditions. These encapsulated conditions completely define the context ("s1") and the predicted outcome ("s2") for individual μ-experiments (postulate H2). A sign is a conjunction of tokens. Individual tokens may be negated (active to inactive, and vice-versa), providing an inhibitory connection. A token retains its time tag when incorporated into a sign.

Postulate T1: **Tokenisation**. Tokenisation is the process by which output from detectors is converted to an internal symbolic form. Such a token symbol may be considered as having a value associated with it that reflects the current (or past) output of the detector. The current token value changes according to the output of the detector.

Postulate S1: **Encapsulation**. Encapsulation is the process by which individual tokens are combined into a single sign. New signs are added to the system during μ-hypothesis differentiation (postulate H6).

Postulates T2 and S2: **Activation**. A token is considered "active" when the detector to which it relates is emitting the output relating to the tokenisation process. Similarly a sign is considered "active" when all its component tokens are

(or were, in the case of time tagged tokens) active, taking into account any negations. Both tokens and signs may be considered as “tests” on the conditions they detect.

3.3.2. Initial Conditions for the Token and Sign Sets

The ethogram will define an initial set of tokens, and ensure they are attached to transducer and detector outputs. A single detector may be associated with several tokens, relating perhaps to different degrees or levels of output. The ethogram will also define any signal processing or transformations to be applied to detector output prior to tokenisation. The initial set of signs will contain one sign for each initial token, unnegated and reflecting the current value of the token. New tokens and signs may be added to the system during the lifespan. Tokens may be defined as active when the state of a transducer changes, either from off to on, or from on to off, or under both conditions. In the experimental conditions described in chapters five and six this effect is inherent in the nature of the environment and simulated transducers. Other environments, real or artificial, may call for specific signal processing to achieve these conditions.

3.3.3. Supporting Evidence for Signs and Tokens

There is a wide diversity of afferent and sensory mechanisms found in nature, and a substantial body of recent research into sensor and transducer systems for artificial animals and robots. This section addresses some of the issues, and presents a sample of sensory strategies to be found in nature. Above all it is clear that sensory sub-systems are far from amorphous, general purpose, elements. Nature abounds with well-documented examples of perceptual mechanisms tuned to the behavioural and learning requirements of their host animal. For instance, Tinbergen (1951, chap. 2) describes how the release strength of the food begging reaction varies in newly hatched herring gull chicks when presented a range of differently coloured model representations of the adult bill. Among many additional carefully observed and documented examples he also reports on the elicitation of the escape response in many species of bird when presented with silhouette profiles of predatory birds, while not reacting to silhouettes of other, non-predatory, species.

Arbib and his colleagues (Liaw and Arbib, 1993; Arbib and Cobas, 1990) have modelled the response of various frog and toad species to the threat posed by large looming objects as possible predators and the opportunity offered by small moving objects as potential prey. Additional neurological evidence that identifiable cells (or structures of cells) respond to external stimuli has been provided by the work of Hubel and Wiesel (1962), who reported that individual cells in the visual cortex become active when highly specific patterns are presented in the visual field of experimental animal subjects. Schölkopf and Mallot (1995) consider the experimental evidence for *place cells*, located in the rat *hypothalamus*, which fire (demonstrate significantly higher rates of electrical activity) when the rat is physically located in specific places.

Tokens, kernels (JCM and ALP) and primitive items (Drescher) are all abstractions from the totality of possible information that will be present at the time the token item is generated. The same is true in nature. The herring gull chick fails to note that the model bill is not a significant feature. The adult bird that the predator silhouette presents no threat - being made of wood and paint. On a different evolutionary path development of the innate releaser indicating this predator danger might be more specific, responding additionally to wing beat patterns, or hovering, swooping or other flight characteristics specific to the predator species. Foner and Maes (1994) point out that many current computer representations of input stimuli only take account of the current situation. This would also appear to be true for the majority of machine learning induction systems. Foner and Maes describe extensions to Drescher's original scheme to allow a one cycle record. This in turn allows extensions to the algorithm to focus attention on phenomena that change. Coincidentally there is also a significant body of evidence for single neurones that demonstrate firing activity specifically with respect to stimulus change.

The evidence for a *Short Term Memory* (STM) phenomena, employed in both JCM and ALP primarily rests with human nonsense syllable recall tasks. The evidence for an activation trace surmised from the apparent ability of various animal species to perform temporal stimulus differentiation. Recent reports implicate the *substantia nigra* brain area as a timing element capable of generating "metronome" like pulses in the millisecond to minute range to other parts of the brain (Highfield,

1996). This is a distinct phenomenon to the daily *circadian rhythm* (Lofts, 1970), which has been demonstrated to influence both physiological and behavioural aspects in a wide variety of species. There is extensive neurophysiological evidence that firing activation can continue after removal of a stimulus at the single neuronal level (an integration effect), though it is not obvious that these phenomena have significant or direct bearing on either the notion of STM or of the activation trace.

The encapsulation of multiple atomic conditions (the tokens) into the single symbolically identified 'sign' (the *sign-gestalt*) allows for an efficient and compact definition of the context-action-consequence triplet representation. Processing transducer and sensor data and hence the derivation of the input token is a critical issue for animat originators. Drescher's *primary items* essentially unambiguously detect a state of the environment that is relevant to the algorithm; the position of the fovea, the location of the simulated hand and so on. By contrast the sensors on the robot used by Mott's ALP system provided highly ambiguous and incomplete information. The same pattern of kernels was generated over a wide range of circumstances. The use of binary representations for light level, for instance, gave ALP little opportunity to determine the true consequences of its actions. In the experiments to be described in chapter six the creation of tokens is tightly coupled to the design of the environment.

3.4. Actions and Reification

The action and reification postulates define the efferent sub-system, which enables the animat to control *actuators* and so directly affect its environment. External actions, those which impinge on the environment, may be monitored by direct observation. Internal actions, such as those which affect the "physiology" of the animat, may only become apparent through measurement or by inference.

3.4.1. The Action Postulates

Definition A0: **Action**. An action is the basic unit of efferent event available to the animat. In the converse process to tokenisation, the animat may convert certain internal symbols into actions that directly impinge upon, and may change, the state of the animat or its environment. In keeping with tradition the terms "action" and

"response" will be used essentially equivalently in this context throughout the thesis[17].

Postulate A1: **Reification**[18]. Reification is the process by which internal symbols are converted into detectable manifestations, for instance physical actions by the animat on the environment via its actuators. Such symbols may be delivered for reification by many routes within the model.

Postulate A2: **Action Cost**. The performance of any action by the animat will be presumed to consume resources otherwise available to the animat. Action costs may be measured in terms of energy expenditure, time taken to completion, or any other units that may be applied consistently within the confines of the ethogram, and which are appropriate to the physical and mechanical design of the animat and its actuators. Action costs are normalised to be 1.0 or greater, where 1.0 is taken as the minimum cost of any of the actions available to the animat.

Postulate A3: **Compound Actions**. Compound actions represent larger sequences of actions, which may be considered as a single tokenised item for reification. They are formed from simple actions (postulate R1) by concatenation. Compound actions formed in this way run to completion once initiated. The cost of a compound action will be taken as the sum of its individual component actions.

3.4.2. Initial Conditions for Actions

The list or vocabulary of actions initially available to the animat is defined in the ethogram. This vocabulary of actions will include all simple and compound actions and their associated costs. New actions may be added to the vocabulary during the lifespan of the animat.

[17] The action as "response" is a S-R behaviourist concept, it is therefore not entirely clear why the term should have been retained by those who did not necessarily regard "actions" as "responses".
[18] (OED) reify v.t. Convert (person, abstract concept mentally) into thing, materialise; hence ~fication n. [f. L *res* thing + -I- + -FY]

3.4.3. Supporting Evidence for an Action Vocabulary

The ethogram may define actions over a wide range of complexity, from simple individual muscle or actuator motions ("molecular" in Tolman's vocabulary, or "characteristic" in McFarland and Sibly's, 1975) to increasingly complex combinations of actions which may be clearly recognised as a behavioural pattern ("molar" in Tolman's and "actions" or "activities" in McFarland and Sibly's). Each animal exhibits a vocabulary of "action patterns", apparently as characteristic of its species as is any physical attribute. The Dynamic Expectancy Model does not divide actions into "appetitive" and "consummatory", as in Tinbergen or Maes' models. In the Dynamic Expectancy Model actions may indeed lead to the satisfaction of a goal (q.v.), but goal satisfaction is rather a property of the goal description, not of any particular action that may precede the satisfying event.

Several detailed studies developing catalogues of essentially unitary behaviour "action patterns" in animals have been undertaken, for instance Shettleworth's work on the Golden Hamster (Shettleworth, 1975) or that of Reynolds' (1976) on the *Rhesus Monkey*. Shettleworth describes 24 mutually exclusive action patterns displayed by hamsters under laboratory conditions. Reynolds' work studied monkeys in a social setting, though in captivity, to prepare an extensive vocabulary of *action patterns*. Action patterns were described as either "postural" (68 distinct actions in 11 groups, including "attack", "threat", "dominance expressions", "submission", "grooming" and "sex") or "vocal", cataloguing the sounds made by his subjects. Reynolds provides comparisons with previous attempts at a terminology and discusses the difficulties in arriving at a uniform and agreed classification.

Mott's ALP used a list of five molecular actions ("<FORW>M", "<BACK>M", "<LEFT>M", "<RIGHT>M" and "<CRY>M"), corresponding to the translational and rotational movements available to the *QMC Mk. IV* robot. It is unclear what role the "<CRY>M" action played in the experimental set-up described. Drescher's system employed 10 molecular actions, four controlling foviation ("eyef", "eyeb", "eyel" and "eyer"), four controlling hand movements ("handf", "handb", "handl" and "handr"), and hand open and close ("grasp" and "ungrasp"). Many

of the simulated and physical robot controllers based on classifier and reinforcement principles define action sets of similar size and complexity.

3.5. Goal Definitions

Goals represent the trigger or cue for the animat to engage in performing outcome directed behaviours.

3.5.1. The Goal Postulates

Definition G0: **Goals**. A goal establishes a condition within the animat causing the animat to select behaviours appropriate to the achievement or "satisfaction" of that goal. Goals are a special condition of a sign; goals are therefore always drawn from the set of available signs.

Postulate G1: **Goal Valence**. From time to time the animat may assert any of the signs available as a goal. Any sign asserted to act as a goal in this way is termed as having *valence* (or be valenced). None, one or many signs may be valenced at any one time. The converse condition, *aversion*, where the animat is required to avoid certain stimulus conditions is considered later (section 7.5).

Postulate G2: **Goal Priority**. Each valenced goal is assigned a positive, non-zero priority. This priority value indicates the relative importance to the animat of achieving this particular goal, in the prevailing context of other behaviours and goals. Goal priority is determined within the innate behavioural component of the ethogram. In the current model only one goal is pursued at any time - the *top-goal*, the goal with the highest priority.

Postulate G3: **Goal Satisfaction**. A valenced goal is deemed "satisfied" once the conditions defined by the goal are encountered, when the sign that defines the goal becomes activated (postulate S2). The priority of a satisfied goal is reduced to zero and it ceases to be valenced. Where goal seeking behaviour is to take the form of sustained maintenance of a goal state, the goal selection process must revalence the goal following each satisfaction event.

Postulate G4: **Goal Extinction**. In a situation where all possible paths to a goal are unavailable, continued attempts to satisfy that goal will eventually become a threat to the continued survival of the animat, by blocking out other behaviours and needlessly consuming resources. Such a goal must be forcibly abandoned. This is the *goal extinction point*. Goal extinction is closely related to the valence break-point postulate (P6).

Postulate G5: **Cathexis**. Cathexis associates a known goal sign with some other sign, following repeated simultaneous appearance. The association grows in magnitude with successive pairings and wanes to extinction should the pairing cease. This mechanism allows created signs to equivalence signs with innate goal properties.

3.5.2. Goals, Starting Conditions and Discussion

Goals are defined within the ethogram, and a mechanism must be defined to enable goals to be asserted whenever an appropriate circumstance arises. Current animat models, based on animal studies, might indicate the appropriateness of goals related to hunger, thirst, internal temperature control, external cleanliness, predator avoidance, location of shelter, mating, and so on (after Tyrrell, 1993). Goal setting and goal satisfaction need not be based on the same detectable phenomena. For instance, food seeking behaviour may be initiated by the detection of lowered blood sugar levels (or by changes in blood sugar controllers, such as insulin). However, due to the delay in the digestive process, were feeding to cease only when these levels were again elevated to a reasonable level the hapless creature would be gorged to bursting point. It has been demonstrated that many cues may be used to terminate feeding behaviour, the action of eating, the taste of sweet but non-nutritious saccharin solution, or by artificial distension of the stomach (by an inflated rubber balloon inserted into the gut). Clearly an overall balance must be achieved between long-term and short-term signals to ensure that behaviour and driving needs are matched.

Goals need not relate to physical requirements, and may be asserted by other mechanisms. Maes (1991) describes “curiosity” as a goal type, related to “exploratory” behaviours. Yet curiosity is rather the description of a process that

involves exploratory or deliberate actions to elicit further information about goals. Such goals may be activated on an arbitrary basis, or specifically to provide additional maturity to a μ-hypothesis, to disambiguate between contradictory μ-hypotheses, or to engage in the process of *play*[19].

3.6. On Policies and Policy Maps

Whenever any goal is valenced (postulate G1) the Dynamic Expectancy Model calls for the animat to construct a Dynamic Policy Map (DPM). As with a *Q*-learning policy map, the DPM allows the animat to select an action based on an estimate of least cost path to the current goal. The DPM is constructed from all the μ-hypotheses available to the system at the time of its construction. Unlike the static policy map of *Q*-learning, commitment to any particular DPM structure and values is not made until the point a goal becomes valenced (G2).

3.6.1. Policy Map Postulates

Definition P0: **Dynamic Policy Map**. The Dynamic Policy Map temporarily assigns a measure of "effectiveness" to every sign known to the animat (the "policy value", q.v.) This effectiveness measure is an estimate of the effort that will need to be expended in traversing from any current situation (as defined and detected by a sign), to the goal sign with the highest given priority (postulate G2). The current DPM is discarded when its goal is satisfied (G3). A new DPM is reconstructed whenever a new top-goal is selected, or when either the set of μ-hypotheses (H5, H6 or H7), or their corroboration measures (H3 and H4) change significantly.

Postulate P1: **Induced Valence**. Any μ-hypothesis whose consequence sign ("s2") is identical to the top-goal sign, or to any sign with valence (postulate G1), induces valence into its context sign ("s1").

---

[19] Play (Dolhinow and Bishop, 1972; Hinde, 1970, pp. 356-359), has been widely observed in animal behaviour, in particular in primates and humans and other mammalian and avian species. Play is not observed in fish, amphibians and invertebrates. Play in animals is most often encountered as incomplete or stylised versions of recognisably adult behaviours, but it is not triggered by normal motivational cues and is without the expected consummatory component. There is a notable suppression of harmful aspects to the normal behaviour manifestation, such as biting. It is also easily interrupted by threat or hunger. Play is often associated with the individual's development in a social context, and as a way of gaining motor skills. It may also have an explicitly exploratory component.

Postulate P2: **Spreading Valence**. Any μ-hypothesis not already valenced, and whose consequence sign ("s2") matches a context sign of another μ-hypothesis that is valenced itself gains valence. Valence is induced (postulate P1) into the context sign, the context sign of the newly valenced μ-hypothesis may now act as a *sub-goal*. Valence may therefore spread throughout the set of μ-hypotheses and signs until all μ-hypotheses have acquired valence, or until no more μ-hypotheses can be reached by this process. The top-goal is defined as having a "valence level" of zero; each level of induced valence increases the valence level by one.

Postulate P3: **Cost Estimate**. The cost estimate for using any action associated with any μ-hypothesis shall be the action cost (postulate A2) divided by the corroboration measure (H3, eqn. 3-1). Thus if the μ-hypothesis has always successfully predicted the consequence its cost estimate (P3) will be equal to the action cost. Where the corroboration measure indicates a less successful rule, the cost estimate rises. Where the μ-hypothesis has always failed the cost estimate would tend to infinity. The reinforcement measure (H4) may be used equivalently in this calculation.

$$\text{cost estimate} \leftarrow \text{cost}(r1) \,/\, p(s2 \mid^{t} s1{+}r1) \qquad \text{(eqn. 3-4)}$$

Postulate P4: **Policy Value**. The *spreading valence* (postulate P2) process creates *policy chains*, indicating one or more paths or chains of actions (extracted from μ-hypotheses implicated in the valenced policy chain) extending between the goal and any sign involved in the DPM. The policy value for any sign that is not the goal and which is involved in the DPM is defined as the sum of individual cost estimates (P3) for each element in the policy chain. In practice the spreading valence method produces a graph or net like structure. Any policy chain shall be defined as comprising the transitions representing the policy cost of lowest overall value between pairs of sign nodes in that chain.

$$\text{Policy value}(s^{n}) \leftarrow \min\left(\sum_{v=0}^{v=n-1} \left(\text{cost}(r1^{v+1}) \,/\, p(s2^{v} \mid^{t} s1^{v+1}{+}r1^{v+1})\right)\right) \qquad \text{(eqn. 3-5)}$$

where *v* is the valence level of each link in the policy chain formed and *n* is the valence level of some sign "s".

Postulate P5: **Action Selection**. Whenever there is a valenced top-goal (and so a DPM) an action may be selected for reification from the μ-hypothesis implicated in the DPM whose context sign is both active (postulate S2) and which has the lowest policy value (P4).

Postulate P6: **Valence Break Point**. Creating a DPM (postulate P2) and selecting an action (P5) establishes within the animat an expectation that the top-goal may be achieved at a certain cost (P4). The model defines a *valence break point* (VBP), typically some multiple of the *policy value* (policy value * n). When actions selected from the DPM fail the policy value rises. Should the policy value exceed that of the previously computed valence break point, goal directed behaviour is suspended, with the animat reverting to exploratory behaviours for a time. During this period the animat may create new μ-hypotheses if the opportunity arises, offering the possibility of a new path chain to the goal. Goal directed behaviour is reinstated with a less demanding valence break point (the policy value is now higher). Goal directed and exploratory behaviours alternate until either the goal is reached, or the goal is finally cancelled by the extinction process (G4). This process mirrors the experimental extinction phenomena repeatably observable in animal experiments (figure 3-1).

3.6.2. Evidence for Chaining

Evidence that animals may form explicit behaviour chains under controlled conditions is described by Blackman (1974). Such chains are created by the experimenter by manipulating the animal in an operant conditioning set-up to elicit some response, say Rx, to achieve a reinforcing reward under some discriminating stimulus situation, say Sx. Following this stage a response, say Ry, is conditioned to Sx, but only in the presence of another discriminating stimulus, Sy. Sx has no inherent reward characteristics, but acts as a *conditioned reinforcer*. Using this method chains of considerable length and complexity have been reported.

$$Sy \rightarrow Ry \rightarrow Sx \rightarrow Rx \rightarrow \text{reward}$$

An independent series of experiments on the *latent extinction* phenomena demonstrates that these behaviour chains may be disrupted, weakened or broken when individual elements of the chain are extinguished (Bower and Hilgard, 1981, describing the work of Stewart and Long, and others.) The ability to construct, and disrupt behaviour chains is not in itself direct confirmation of induced valence, but is important supporting evidence. Experience from animal training (Bower and Hilgard, 1981, p. 179) suggests that the chain need not be built up backwards from the primary source of reinforcement, but may also be built forwards, or by inserting operant elements into existing shorter chains.

3.6.3. Evidence for Goal Suspension and Extinction

Figure 3-1 shows stylised cumulative records (from Blackman, 1974, p.67, after Reynolds) derived from Skinner box experiments under various operant conditioning reinforcement schedules. In the *fixed ratio* (FR) schedule "reward" is delivered to the animal after a fixed number of "responses". In the *variable ratio* (VR) schedule "reward" is delivered after a random number of "responses". In the *variable interval* (VI) schedule "reward" is delivered at randomly varying intervals, independently of actions by the animal. Similarly, the *fixed interval* (FI) schedule delivers "reward" after a fixed interval of time, again independently of "responses" by the subject. All these schedules are applied to animals that have previously been conditioned to operate the Skinner box apparatus on a regular reward schedule.

The slope of the curve indicates the rate of the learned response (each response causes an upwards increment in the trace), downward "tick" marks indicate individual reinforcing reward events. Note the characteristic stepped form of the curve in the *extinction* phase of the experiments following the cessation of reward events. The stepped form reflects the changing relationship between two forms of activity during the extinction phase, shortening periods when responses are made, and lengthening periods when no responses are made. In time the learned response is apparently completely eradicated. This extinction process is a highly repeatable phenomenon, and has been widely reported under both classical and operant conditioning regimes. Experimental regimes also indicate a secondary process of

spontaneous recovery, in which the previously extinguished effect re-appears, albeit in a weakened form, after a period of rest.

The Dynamic Expectancy Model emulates the shape of the extinction curve by the combined effects of the reinforcement (H4), valence break point (P6) and goal extinction (G4) postulates. Specific details of how these interact in the implemented model, and experimental analysis of the effects are described later. Extinction curves of the type shown in figure 3-1 indicate the manner in which an animat may abandon use of individual μ-hypotheses that prove ineffective. The reinforcement schedules themselves may yet reveal much about how μ-hypotheses may be created and managed in an animat designed with biological plausibility in mind.

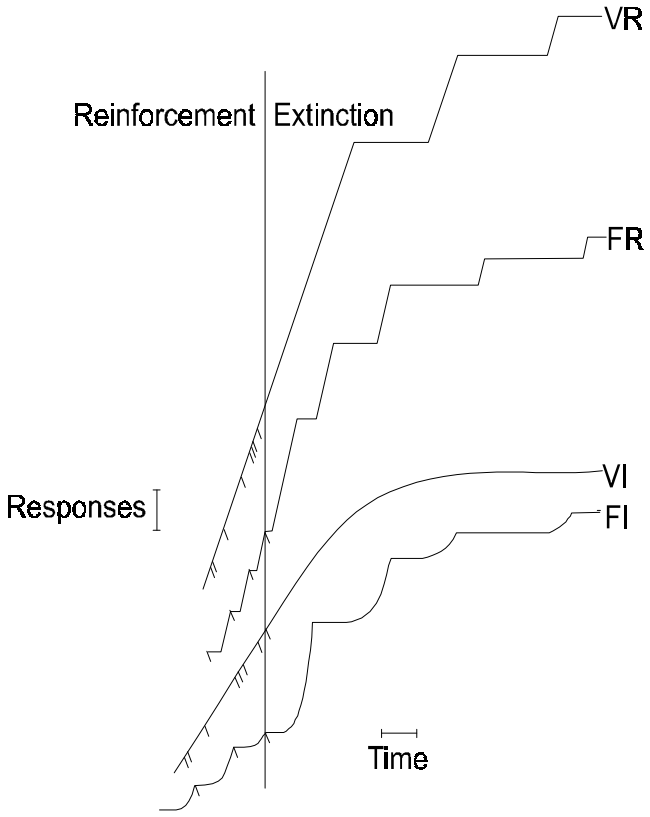

Figure 3-1: Extinction Curves Under Various Schedules

3.6.4. Comparison to *Q*-learning

The Dynamic Expectancy Model is based on a different set of fundamental premises to that of the reinforcement and *Q*-learning strategies of Sutton and Watkins. Watkins (1989, p.16) summarises the situation for *Q*-learning in three position statements: (1) that the capacity for maximally efficient performance is valuable; (2) that exploration is cheap; and (3) that the time taken to learn a behaviour is short compared to the period of time during which it will be used. Statement (1) is hardly in contention. Statements (2) and (3) indicate that the ultimate level of performance is inherently more important than the time taken to achieve it. "Optimality" is thus defined as maximising reward acquisition over an extended time period. Learning in the Dynamic Expectancy Model aims to provide the animat with the best path to achieve goal (reward) states as they become indicated, given the current level of knowledge. It may be that as the animat becomes more experienced the quality of that path might be expected to converge to some acceptable notion of "optimal"[20] behaviour. This would be the case, as discussed under the FDMSSE conditions considered earlier, except for the competing requirement that the animat continue to explore while any phenomena remain unpredicted, an innate drive to continuously augment and refine its state of knowledge.

3.7. Innate Behaviour Patterns

Innate behaviour patterns provide a grounding for intelligence. In the Dynamic Expectancy Model innate behaviours serve three distinct roles. First they provide the animat with sufficient behaviour to survive in its environment from *parturition*, before any learning. These behaviours imbue the animat with strategies to react to life threatening events, where learning would represent too high a risk for failure on the initial instances; predator avoidance for instance. Second to select and set goal priorities. Most goal directed behaviour serves basic physiological requirements. Innate behaviour detects conditions indicating those requirements and establishes them as goals. Third to provide a level of background behaviour to

[20]Optimality, like beauty, is in the eye of the beholder. The *Q*-learner may regard the shortest path between current state and reward state as the optimal path. A hungry predator waiting beside this path may agree.

ensure the animat is appropriately tasked whenever neither the primary nor secondary roles are activated. It may be appropriate that the animat enters a state of hibernation, torpor or sleep, a strategy that may conserve energy or serve other physiological functions. The animat may also use these periods to perform exploratory actions, thereby triggering μ-hypothesis creating postulates, and performing acts that corroborate existing μ-hypotheses. It is a consequence of the Dynamic Expectancy Model postulates that learning may take place in the absence of explicit reinforcement. Several strategies for this exploration may be applicable.

Definition B0: **Behaviours**. Behaviours are unlearned activities inherent within the system. Behaviours give rise to actions (postulate A0) in response to circumstances detectable by the animat. They are defined prior to parturition as part of the ethogram. There is no limit to the complexity (or simplicity) of innate behaviour. An animat might be solely dependent on innate behaviours, with no learning component.

Postulate B1: **Behaviour Priority**. Each behaviour within the animat is assigned a priority relative to all the other behaviours. This priority is defined by the ethogram. The action (postulate A0) associated with the behaviour of highest priority is selected for reification (A1).

Postulate B2: **Primary Behaviours**. Primary behaviours define the vocabulary of behaviour patterns available to the animat at parturition. These behaviours provide a repertoire of activities enabling the animat to survive in its environment until learning processes may provide more effective behaviours.

Postulate B3: **Goal Setting Behaviours**. The ethogram defines the conditions under which the animat will convert to goal seeking behaviour. Once a goal is set the animat is obliged to pursue that goal while there is no primary behaviour of higher priority. Where no behaviour can be selected from the DPM, the animat selects the behaviour of highest priority that is available. Behaviour selection and reification (A1) from the DPM resumes once there is any match between the set of active signs (S2) and the current DPM (P5).

Interruption of goal directed behaviour by a higher priority innate behaviour may draw the animat away from its top priority goal. For instance, goal directed nourishment seeking behaviour may be interrupted by high priority predator avoidance activity. Once the threat is passed goal directed behaviour will be resumed, although the animat's perceived "place" in the DPM graph will have shifted as a result of the intervening behaviour. The structure and corroboration of the DPM may have changed, and it must be re-evaluated as behaviour reverts to the goal directed form. Where goal seeking takes the form of a sustained maintenance of the selected goal state, the selection process must reassert the required goal each time it is satisfied.

Postulate B4: **Default (exploratory) Behaviours**. Default Behaviours provide a set of behaviours to be pursued by the animat whenever neither a primary nor goal setting behaviour is in force. Typically these default behaviours will take the form of exploratory actions. Exploratory actions may be either random (*trial and error*), or represent a specific exploration strategy. Selection of this strategy will impact the rate and order in which the μ-hypothesis creation processes occur (H5). Default behaviours have a priority lower than any of the primary (B2) or goal setting (B3) behaviours. The provision of default behaviours is mandatory within the ethogram.

3.7.1. Balancing Innate and Learned Behaviour

The balance between innate and learned behaviour varies widely throughout both nature and the study of artificial animats. Action selection models, such as those of Brooks, Chapman and Agre, Maes, and Tyrrell, place full emphasis on the provision of pre-programmed behavioural activity. Behaviours are selected to give the animat appropriate responses to its environment, and as a consequence animat behaviour may appear "intelligent" by virtue of this applicability. In this case the originator imbues the animat with a mechanism to determine which needs are required, and a mechanism to balance between them. Within its repertoire of innate behaviours a simulated animal may manage its requirements for nourishment and water, for warmth, for shelter, predator evasion and the need to procreate.

Similarly a robot may be programmed to partition its activities into different, and mostly mutually exclusive, behaviours - collecting soda-cans, environmental

mapping, avoiding unexpected obstacles, seeking its recharging point and replenishing its batteries. Each robot may incorporate these, and other tasks, whose usefulness and complexity are limited primarily by the imagination, patience and programming skills of the robot designer. Recall that μ-hypotheses may themselves be defined in the ethogram, consequently the Dynamic Expectancy Model does not imply that all goal seeking behaviour must be learned.

At the other end of the scale many adaptive learning models adopt a *tabula rasa* approach. With little or no predefined coherent behaviour, they rely instead on a (pre-defined) learning mechanism to accumulate sufficient information about the environment to eventually create coherent and appropriate overt behaviour. Reinforcement and *Q*-learning schemes fall into this category, as does Drescher's schema system. Initially actions are selected at random, under a *trial and error* regimen and internal structures built or existing structures populated. With the application of sufficient trials purposive behaviour may be generated from the structures and information accumulated.

Mott's ALP was essentially initially a *tabula rasa* system, but a small number of low-level robot reflexes were provided. To prevent the robot becoming physically trapped into corners a reflexive backoff mechanism was pre-coded into the robot control-level controller. This is a recurring problem for mobile robot constructors, exacerbated in this instance due to the physical layout of the robot used, a square outline with differentially powered wheels forward of the centre-line. For this reason many mobile robots are designed with a circular, or at least rounded "floor-plan", with their drive wheels placed symmetrically about the centre-line. A second low-level innate reflex was found to be necessary to suppress the backoff reflex when the robot was at the charging point. This "discriminating push" reflex prevented contact with the charger being broken, ensuring that effective electrical contact was maintained between the robot's charger contact plates and the sprung base station charger contacts throughout the recharging period.

3.8. Advances Introduced by the Dynamic Expectancy Model

Cursory inspection of the *Dynamic Expectancy Model* postulates H3 and H4 might suggest that this is a conventional reinforcement model of learning. Procedures

(encapsulated by equations 3-1, 3-2 and 3-3) by which reinforcing events strengthen or weaken disposition of the animat to adopt one behavioural option over another are similar to those of other well-established reinforcement methods. The source of the reinforcement is, however, radically different. In the Dynamic Expectancy Model the reinforcement signal is internally generated by the setting and subsequent verification of a prediction. In previous reinforcement systems the reward signal must be received from the external environment before any learning could occur. In the new model a valid reinforcement signal is generated whenever a behaviour choice is exercised and a μ-experiment activated, so that the processes of behaviour may now be largely disassociated from those of learning.

It will be demonstrated later that this new method allows for substantially improved learning rates over conventional reinforcement learning techniques (section 6.2). It is quite clear that learning triggered by external reinforcing reward is also a valid effect, and commonly observed in animals. While this thesis primarily explores the effects of internally generated reward, it will be demonstrated (section 7.4) that additional performance benefits may accrue to the animat when internal expectancy and external reward signals are combined.

The Dynamic Policy Map arises from the fundamental disassociation of the learning and (goal-seeking) behavioural processes. In the static policy map of, say, the *Q*-learning algorithm, each sensory state becomes increasingly permanently attached to a particular action relative to a fixed goal. While this may bring advantages in enhanced reaction times following the learning phase, it leads to an inflexible reaction to the changing needs of the animat with time and varying goals. The Dynamic Expectancy reinforcement method of learning allows the construction of a policy map only when it is required, and relative to the specific needs of the animat at the time of construction. μ-Hypotheses become "committed" to a particular goal only while that goal has the highest priority, and will be reallocated whenever the goals of the animat change. An example of this dynamic map construction will be given in section 4.9.3.

By generating the policy map dynamically in this way the advantage of the reactive response to active signs inherent in the static policy map is retained. By not committing any individual μ-hypothesis to any particular goal or reward during the

learning process the Dynamic Policy Map may be reconstructed to provide a reactive policy relative to the current goal, even where the goal has not previously been implicated in the learning process.

By integrating expectancy learning with an action selection based model of behaviour a way of selecting goals is made possible. This combination of techniques also provides a way of defining innate, reactive stimulus-response behaviours. These innate behaviours provide the animat with a mechanism with which to react in a manner to allow survival while the individual learns the skills required to behave ever more appropriately in its environment.