Chapter 8

8. Discussion and Conclusions

8.1. Reactive or Cognitive?

The initial problems remain. Is behaviour in animals and animats primarily or wholly according to responses mediated by the immediate reaction to impinging stimuli? Is learning simply a matter of strengthening or weakening the connections between stimulus and response, as the reactive or situated agent behaviourists would have us believe? Or is behaviour primarily instigated by "goals", internal states of the animat set and satisfied according to the physiological needs of the animat, with the processes of the animat selecting actions to pursue those goals?

These questions have been hotly debated for nearly a century, with a mountain of evidence accumulated for both viewpoints. Brooks (1991b) has argued (and many before him), much of what we observe in animal and human behaviour can be perfectly adequately explained with a purely stimulus-response analysis. Yet from the time of Tolman (1932) psychologists have argued that reactive behaviourism is wholly inadequate to explain the behavioural abilities of the human species and, as demonstrated through ingenious experiment, to explain all the behavioural abilities of animals.

8.2. Expectancy Model as "Missing Link" in Learning Theory

The Dynamic Expectancy Model may be thought of as the "missing link" between pure S-R behaviourism and the "cognitive", goal based, approach. While the Dynamic Policy Map is created by a goal driven process, utilising the three part representation of the μ-hypothesis, a purely cognitive notion, immediate behaviour is selected only on the basis of the current stimulus set, and so may be thought of as purely reactive. In many experimental designs the two may appear almost indistinguishable from one another. A similar distinction has been developed in the

idea of universal planning, which is considered in more detail later in this chapter. Critically, and in keeping with the observation that reward is most effective if applied immediately following an event, reinforcement is still applied directly to the main unit of learning, the μ-hypothesis, immediately the outcome (of the prediction) is known. The adaptive component of the learning process is pure reinforcement; behaviour due to the combination of these units to produce goal seeking behaviour by the *spreading activation* process. Direct reinforcement relative to a known system "motivation" is not excluded, as demonstrated by the *valence level pre-bias* experiments. There is also no restriction to the re-ordering or strengthening of elements of the Behaviour List $\mathcal{B}$ in a manner entirely consistent with a pure S-R behaviourist reinforcement regime.

Given the obvious diversity of both physical and behavioural characteristics across all the species of the animal kingdom, it would appear idle to suggest that there would not be a similar diversity of behavioural and learning strategies. Some animals with simple behavioural strategies may employ no adaptive ability, or limited learning strategies. In others the number and complexity of these strategies increase, manifest as improved behavioural ability. Razran (1971, p. 252) has proposed an "evolutionary ladder of reactions", which argues for a correlation between an animal's place on the evolutionary scale with the appearance of experimental evidence for various learning strategies at the different levels. In this context adoption of different and varied reinforcement strategies, and similar strategies to varying extents, by different species seems inevitable.

8.2.1. Types of Reinforcer

The conventional view of a *reinforcer* is related to underlying biological needs, such as "*food, water or sexual contact for appropriately deprived individuals*" (Bower and Hilgard, 1981, p. 268). It is exactly these needs that can repeatedly be demonstrated as the motivations or drives to initiate and sustain behaviour. It makes design sense to learn behaviours relating directly to those aspects that will be most germane to the everyday existence of the animat. Such *primary reinforcers* may be easily identified and categorised into phenomena that do, and those which do not, act to modify behaviour. In SRS/E, with the valence level pre-

bias (VLPB) option enabled, any sign placed on the Goal List will subsequently adopt the role of a primary reinforcer.

It is clear that phenomena other than direct biological need can act as a learning reinforcer. Such *secondary reinforcers* may include "*money, praise, social approval, attention, dominance and the spoken exclamation "good"*" (Bower and Hilgard, 1981, p. 268). At a level below even the primary reinforcers, notions of "pleasure" and "pain" appear to "pre-classify" stimuli and sensations into desirable phenomena, to be sought and undesirable phenomena, to be avoided. The existence of specific nerve types to detect "painful" stimuli would indicate that this is a very primitive mechanism, one it is easy to argue will have a very immediate impact on the survival rate of an organism. "Pleasure", on the other hand, seems to be associated with a much higher level of neural organisation. In this context the application of expectation satisfaction appears as a bridging reinforcer. Expectation satisfaction is neither a primary reinforcer - it serves no direct biological need, nor a secondary reinforcer - as it does not require a social infrastructure implicit in the list of secondary reinforcers.

8.3. Relationship to Policy Maps and Universal Plans

A feature of the *Dynamic Policy Map* is that it indicates the most appropriate action to take in the specific set of circumstances defined by the goal being sought and by the prevailing sensory pattern. In SRS/E this pattern may include elements from the trace of past sensations. In this respect the action selection mechanism has many similarities to the *policy map* described for reinforcement and *Q*-learning procedures. These procedures suffer in comparison to the DPM when the goal definition changes, or the path to the goal becomes blocked or radically altered. Schoppers (1987, 1989, 1995) develops the notion of *universal planning* that addresses the plan/react issue from a different direction.

In Schoppers' system a conventional planner builds a problem-solution path using goal reduction operators. The resulting structure is converted into a decision tree. This may be traversed for each current situation to determine the action appropriate to the prevailing conditions defined by a set of known and predetermined predicate tests, a *cache* of pre-formulated step solutions. The

reactive nature of the universal plan overcomes a form of brittleness inherent in conventional planning, where failure of any stage during execution causes failure of the plan as a whole. Universal plans react to successes and failures in activity without recourse to additional computationally expensive replanning.

Ginsberg (1989) argues against the universal plan as a useful approach. He argues that the size of the cache will grow exponentially with the number of sensors, that there will be only a minor computational cost saving, and that this will be at the expense of greater storage requirements. Ginsberg's exponential growth argument is based on the notion that all sensors are independent, and that each sensor may be connected to every action. He further argues that, unless the "universal plan" covers all eventualities it should properly be referred to as an *approximate universal plan*.

Strict application of the exponential complexity argument is specious. The world is clearly non-uniform. Were the world "uniform" then it would make no difference which action was taken under what circumstances, and such is palpably not the case. All associationist, behaviourist and cognitive models are based on the exploitation of this non-uniformity. Rivest and Schapire (1990) have presented an algorithm to detect and utilise equivalence in detectable conditions. Using this algorithm the 10^{19} states of the sometime popular children's toy the *Rubik's Cube* may be reduced to 54 conditions. Yet it may be that important conditions in the environment are poorly distinguishable, either because they are in some true sense similar, or because the sensory capabilities used to differentiate between them are ineffective. Under these conditions the behavioural (and learning) mechanisms will be obliged to incorporate a broader spectrum of sensations to disambiguate between candidate options.

If we view the evolution of species as nature's "universal plan generator" (as made manifest in an individual's ethogram), it becomes clear that these exponential complexity pre-conditions relating to sensors do not hold. As discussed in an earlier section, nature apparently tailors and tunes otherwise undifferentiated sensory apparatus to each task. Tinbergen's birds responded quite specifically to certain "predator" silhouettes, but were apparently oblivious to other shapes. SRS/E and other like systems may take advantage from similarly tuned sensory

apparatus, but even without this advantage will seek to identify those combinations of sensations that are significant, and ignore the remainder. In summary there is no need for sensory apparatus to be uniform or homogeneous.

Classical AI planning systems have two potential advantages over reactive and policy based approaches. First, they are (or should be) incorporated into formally correct search procedures. More significantly this implies that the operators defined must themselves be correct; that is achieve the outcome they promise, under the conditions they promise them. Second, the classical planner may take different actions based solely on its current position in its internal solution path, although the incoming sensor vector is identical. The current detectable conditions are used for confirmation, or not at all. Purely reactive systems based on the current sensor vector do not have this advantage. SRS/E addresses this problem by the use of activation traces and recency values. Other approaches may allow recirculation of sensory data (for instance, Becker's proposal to re-circulate kernels into STM,) or some other method for the explicit recording of past events into the representation.

However, classical AI planning can lead to a form of brittleness. If the operators are not correct the solution path generated will not be correct. Advantage gained from the correctness of the search procedure is compromised. SRS/E operators, the μ-hypotheses, are, by their nature, only an estimate of the described transition. The Dynamic Policy Map allows the SRS/E algorithm to select actions on the basis of combined probabilities, as manifest in the cost estimation procedures, and then to update its confidence in individual μ-hypotheses on the basis of the outcome. It is particularly robust in the face of unexpected outcomes caused, among other reasons, by faulty or unconfirmed μ-hypotheses. It takes advantage of serendipitous transitions forward to the goal where the cost estimate unexpected falls; and may continue along some other route to recover from a failure to traverse the expected path.

In a wide range of circumstances speed of response is the critical issue in behaviour. The tardy prey, absorbed in careful planning of its escape, might expect no quarter from the stooping hawk. Perhaps predictably, Schoppers (1989) in his reply to Ginsberg argues in favour of the increased space utilisation for the cache to achieve responsiveness. Given the incompleteness of most behavioural

repertoires, and of the scope of the current generation of formal planners, "universal plan" may indeed be something of a misnomer.

8.4. One-Shot Learning Phenomena

The SRS/E model clearly demonstrates the *one-shot learning* phenomena. As soon as the μ-hypotheses is created the animat has a possible path between the two points in the "cognitive" map represented by the signs "s1" and "s2" embedded in a μ-hypothesis. An effective μ-hypothesis becomes rapidly adopted as the path of choice, and the animat will appear to learn quickly, possibly as a result of a single trial. Because its outcome is successfully predicted, discovery of an effective solution also has the effect of suppressing further learning activity related to the sign "s2". If, as is more likely, the new μ-hypothesis fails to encapsulate all the conditions necessary for a perfect prediction, further learning may occur at each instance of an imperfect prediction. At some point it may be that there are sufficient imperfect μ-hypotheses to ensure that every instance of "s2" is predicted, and learning for this restricted sub-domain will cease, at least temporarily.

This procedure may serve to explain the conundrum (described by Bower and Hilgard, 1981, p. 341) of why a rapidly learned path is quickly extinguished, yet one that is learned over an extended period takes longer to disappear. Individual μ-hypotheses are (in SRS/E at least) extinguished at an essentially equal rate, on the basis of activations, not elapsed time. Where one-shot learning has taken place, a single μ-hypothesis is available to reach the solution while the goal is asserted. No further μ-hypotheses being created as none are required. The observed extinction time is therefore equivalent to that for a single μ-hypothesis. Where several such alternative, albeit imperfect, μ-hypotheses exist, more than one path will be available through the Dynamic Policy Map. As each path fails, another will be selected from the recomputed DPM. The animat will continually swap between the alternatives as the estimated policy cost shifts (at a rate determined by the parameters previously discussed) due to prediction failures. Eventually one, then another and finally all the different paths are extinguished and the goal is finally abandoned as unachievable in the normal way.

Overall time to extinction, as measured by the count of actions ascribed to pursuing the goal, is then (in the SRS/E algorithm at least) a function of the number of alternative paths through the DPM. Alternative paths arise through imperfect μ-hypothesis formulation, which extends learning time. Therefore, extended learning times lead to extended extinction times. Careful examination of results from extinction experiments (section 6.5) reveal this effect, which is particularly apparent in the *dual-path extinction* procedures (figure 6-17).

Taken to a natural conclusion, SRS/E attempts to build a hypothesis about every sign it might detect, and also to predict every occurrence of those signs. Under certain circumstances these conditions can hold true, for instance those described by some *Markov Decision Processes* (MDP) worlds. In the finite and deterministic (FDMSSE) environment the SRS/E algorithm will stabilise with a μ-hypothesis to predict every sign and for every appearance of each possible sign.

8.5. Expectancy Theory and XBL - a Proposal

The development of expectation based learning directly impacts one of the long standing conundrums associated with machine learning; how to make learning truly autonomous. Autonomous learning means that the animat or learning program can learn without any form of external supervision or guidance as to what represents a "good" or "bad" choice. In the case of the novel Dynamic Expectancy Model described in this thesis, and tested in the form of the SRS/E algorithm and implementation, a reinforcement signal is generated internally from successful and failed predictions.

Generally machine learning algorithms fall into two categories, supervised and unsupervised learning. In the former category a teacher is on hand to indicate to the system the appropriateness of its actions and so provide the feedback to guide the learning mechanism. In the latter case information about the task to be learned has been embedded in the code. Buchanan, Smith and Johnson (1979) refer to this component as the *critic*. The critic compares the outcome of the *performance element*, responsible for the overt (and possibly faulty behaviour) with the predefined desired behaviour and supplies an error or difference signal to a *learning element*, which modifies the performance element accordingly. Their

model of machine learning is a general one, but the form in which each of the elements appears and the nature of the signals passed between them is particularly diverse.

Expectation Based Learning (XBL)[30], based on the principles laid down for the Dynamic Expectancy Model, at last releases the *etho-engineer*[31] from the obligation, but not the option, to specify goal or purpose related criteria for the learning element. Evaluation of an SRS/E μ-hypothesis on the basis of its predictive ability forms a measure of the effectiveness of that μ-hypothesis. Its usefulness is a separate issue, related to the degree to which it enables the performance element to pursue some pre-defined or otherwise generated purpose. The *valence level pre-bias* (VLBP) experiment demonstrates that when learning and performance are indeed linked, both may be advantaged.

Drescher (1991) suggests the term "Schema Based Learning" be adopted as appropriate to the class of intermediate level cognitive models. Notwithstanding the importance of the tri-partite representation adopted by SRS/E, ALP and JCM, it, however, does not align directly with the notion of expectancy. The satisfaction of an expectancy is not tied to this particular representational formulation. It is possible that the notion of an expectation and its subsequent satisfaction may prove to be applicable to a wide range of other otherwise quite conventional structures already employed in the fields of Artificial Intelligence, Machine Learning and Adaptive Behaviour research.

[30] XBL, rather than EBL, as this term is already in widespread use ("*Explanation Based Learning*", Minton *et al*, 1990)
[31] One who engineers ethograms - for want of a more apposite term