An Action Selection Calculus

Mark Witkowski

Department of Computing
Imperial College London
180 Queen's Gate, SW7 2AZ
United Kingdom
m.witkowski@imperial.ac.uk
Tel: +44 (0)20 7954 8423
Fax: +44 (0)20 7581 8024

An Action Selection Calculus

(An Action Selection Calculus)

Abstract

This paper describes a unifying framework for five highly influential but disparate theories of natural learning and behavioral action selection. These theories are normally considered independently, with their own experimental procedures and results. The framework presented builds on a structure of connection types, propagation rules and learning rules, which are used in combination to integrate results from each theory into a whole. These connection types and rules form the Action Selection Calculus. The Calculus will be used to discuss the areas of genuine difference between the factor theories and to identify areas where there is overlap and where apparently disparate findings have a common source. The discussion is illustrated with exemplar experimental procedures. The paper focuses on predictive or anticipatory properties inherent in these action selection and learning theories, and uses the Dynamic Expectancy Model and its computer implementation SRS/E as a mechanism to conduct this discussion.

Action Selection, Learning Theory, Behaviorism, Conditioning, Expectancy Model, Anticipation

1 Introduction

This paper reviews and then proposes a unifying approach to five highly influential classic theories of animal behavior, action selection and natural learning. These are Stimulus-Response (S-R) behaviorism, the associationist model, classical (Pavlovian) conditioning, operant (or instrumental) conditioning and Sign-learning or *means-ends* (pre-cognitive) models. Collectively, these will be referred to as the *five factor theories*. During the 20th century, each theory attracted strong proponents and equally strong opponents and each was dominant for a time. These theoretical positions are placed in their historical context in this paper. The legacy established by these theories endures and each retains a strong influence on elements of contemporary thinking.

Each theoretical position is supported by (typically large numbers of) detailed and fully repeatable experiments. None of these theories completely explains the full range of observable behaviors for any animal species and none was able to gain an overall dominance of the others. It is argued here that is useful to consider each as a partial theory. Each explaining – often in remarkable detail – some limited aspect of the complete behavioral and learning repertoire of each animal. And that to better understand the overall behavior of an animal each of the theories must be placed into a single unifying context.

However, the fact remains that the consequences of each of these theoretical positions can be demonstrated in a single animal – though not all animal species will necessarily demonstrate every attribute. Each is made manifest in the animal according to the experimental procedures and circumstances to which it is subjected. In developing the unifying approach, examples will be drawn widely from the animal and *animat* (artificial or simulated animal) research domains and these will be balanced between the contemporary and historical context. Each of the five factor theories is characterized by the underlying assumption that observable and measurable behavior results from sensations arising from the interaction between the general environment of the organism (including its body) and its sense organs. The issue under debate was, and remains, the principles by which this interaction is to be characterized. In itself, overt behavior gives little immediate indication of which,

indeed, if any, of these theories best describes the internal action selection mechanism that gives rise to the observable behavior.

The task here, then, is to provide a minimal description of the principles underlying the mechanisms involved that recognizes natural diversity, yet covers the range of phenomena observed. The approach here is reductive, abstracting from the mass of experimental detail to reveal a broader unifying context. Starting with these five theoretical positions, a semi-formal system is developed consisting of three connection types, five propagation rules for behavioral action selection and four learning rules. This is the Action Selection Calculus. The purpose of this endeavor is to isolate the principal properties of the five factor theories in relation to the behavioral action selection problem and to construct an overall framework with which to consider the role of particular instances of action selection behavior within the *complete behavioral repertoire* of an individual animal or species.

The Action Selection Calculus developed here provides a framework for describing behavioral action selection and learning typically found in non-human animals. While the five factor theories can be applied to certain aspects of human behavior and may form a significant substrate to low-level human behavior, it is clear that even taken together they substantially fail to explain the diversity and richness of the human behavioral repertoire. Outstanding progress has been made in the *cognitive sciences* (e.g. Boden, 2006 for a broad view) in the years since the five factor theories were propounded and the direct influence of these theories on our understanding of human behavior is now very restricted. Nevertheless, the continuing contribution of these theoretical positions and the detailed animal experimentation that underpins them should not be undervalued. The Action Selection Calculus both revives and integrates these theories into an operational model of animal action selection.

Equally, the Artificial Intelligence *Machine Learning* community has contributed greatly to our understanding of the basic principles of human and animal learning, notably in the areas of *Reinforcement Learning* (Kaelbling, Littman and Moore, 1996; Sutton and Barto, 1998) and *Artificial Neural Networks* (Bishop, 1995). See Langley (1996) for a broader review of Machine Learning techniques.

The development of this Action Selection Calculus stands as a theoretically and computationally motivated experiment to reify and formalize these different (and sometimes competing) models of learning and action selection. It is not the purpose of this paper to reject or deny any of the factor theories, but rather to postulate a mechanism by which they might all be manifest in the one animal or animat, each (one can only presume) serving a valuable function within the creature as a whole, so as to enhance its overall chances of survival.

The Calculus does not presuppose any particular class or species of animal, yet the elements of the Calculus cannot be uniformly applied across the animal kingdom, as it is clear that the underlying factor theories cannot be applied uniformly either. Razran (1971) relates the occurrence of different behavioral and learning strategies to position in the phylogenic scale, pointing out that S-R mechanisms are ubiquitous and that almost all animals with even a rudimentary nervous system may be classically conditioned. Operant conditioning and means-ends behavior are both clearly manifest in mammalian species. Opinions vary widely as to the relative importance that should be placed on each of the factor approaches, most standard texts (e.g. Bower and Hilgard, 1981; Hergenhahn and Olson, 2005) treat them separately without any significant attempt at integration. This is now addressed by this paper. Significant pieces of work that attempt an integration of two or more of the factors are identified in section 2.

The rat might well serve as an exemplar of an individual animal that clearly demonstrates all the attributes of the Calculus, with simple S-R reflexes, a rich innate repertoire of individual and social behaviors, which has been used extensively in the laboratory to demonstrate both classical and operant conditioning, as well as means-ends behaviors. The simulation experiments described later in the paper are abstracted animat style emulations of the type conducted with rats in the laboratory. Barnett (1973), for instance, provides a comprehensive account encompassing both the natural behavior of the rat and its behavior within the confines of various laboratory procedures.

The Action Selection Calculus notation developed here highlights certain aspects of the problem. It is by no means the only possible notational approach, each of which predisposes its own emphasis. Emphasis is placed on the role of anticipation and prediction, both as a guiding principle by which

organisms may express their fitness for the environment they inhabit and as a mechanism to achieve learning and behavioral action selection – the *anticipatory stance*.

All the connection types in the Calculus are between *Signs* (stimuli), detectable sensory conditions, and *Actions* (responses), behaviors that may be expressed by the animal or animat. There are many possible combinations of Signs and Actions. The three selected here each encapsulate an anticipatory or predictive role, one implicitly, and two making and using explicit predictions.

This paper also identifies where the factor theory mechanisms clearly differ and where they *apparently* differ, but can be explained as manifestations of a single type of mechanism and how these differences may be resolved into a single structured framework. It sheds light on why it has traditionally been so hard to resolve between responsive and goal directed action selection approaches and considers the development of motivation across this boundary (e.g. Brooks 1991a; 1991b; Kirsh, 1991). The role of reinforcement and predictive anticipation in the learning process is also investigated and the properties of these two approaches compared. Perhaps surprisingly, this analysis reveals that two of the approaches, namely, operant conditioning and means ends – long regarded as diametrically opposed – may be expressed by a common mechanism.

The Action Selection Calculus is an abstraction of the 20 *Dynamic Expectancy Model* (DEM) *postulates* (Witkowski, 2003). These are a detailed specification for the action selection and learning properties of the Dynamic Expectance Model. The 20 postulates also form a high-level specification for the actual (C++) computer implementation of the DEM, *SRS/E*. The SRS/E program has been used to conduct a substantial battery of simulation tests (Witkowski, 1997), mimicking or replicating actual experimental procedures on animals. A small selection will be presented to illustrate points made in this paper. Aspects of SRS/E will be used to reify specific attributes of the Action Selection Calculus.

Section 2 provides a thumbnail sketch of each of the five factor theories. This paper is not intended as a detailed review of the evidence supporting each theory, each of which is already a digest of many exemplar and corroborating experimental procedures. Comprehensive descriptions of the five theories and their evidential support can be found in any textbook of natural learning theory (e.g. Bower and Hilgard, 1981; Hergenhahn and Olson, 2005). Section 3 considers the sensory and motor interface between animal and its environment and how issues of behavioral motivation might be addressed.

Sections 4, 5 and 6 respectively build the arguments for the structural, behavioral and learning components of the combined approach. Section 7 reconstructs the major factor theories in the light of these component parts and emphasizes the role of the action selection *policy maps*[1], which may be either static or dynamic[2]. Section 8 describes an arbitration mechanism between these policy maps, leading to final action expression. Section 9 presents some illustrative experiments to show various properties of the Action Selection Calculus. Section 10 presents some discussion of the topics raised in this paper. Appendix one tabulates and defines the notation commonly used throughout. Appendix two describes aspects of the SRS/E design and implementation.

2 Theories of Action Selection Behavior and Learning in Animals

We continue with the view that animal behavioral action selection is properly described by the direct or indirect interaction of sensed conditions, Sign-stimuli (S) and response, action or behavior (R) initiators. This section will briefly outline five major theoretical positions initially formulated in the first half of the 20th century relating to animal behavior and learning and summarizes the continuing influence they have exerted on subsequent thinking. In particular it will focus on those issues relating to action selection, which will be considered in detail later. In each case the basic principles will be considered from a historical perspective, but each will be illustrated with examples of research drawn from the recent and contemporary corpus.

2.1 The Stimulus-Response (S-R) Behaviorist Approach

It has been a long established and widely held truism that much of the behavior observed in natural animals can be described in terms of actions initiated by the current conditions in which the animal finds itself. This approach has a long tradition in the form of *stimulus-response (S-R) behaviorism*[3]. Although this was proposed in its modern form over a century ago (Thorndike, 1898), it still continues to find proponents, for instance in the behavior based models of Maes (1991), the reactive or situated models of Agre (1995) and Bryson (2000), and was a position vigorously upheld by Brooks (1991a, 1991b) in his “intelligence without reason and representation” arguments.

All argue that the majority of observed and apparently intelligent behavior may be ascribed to an innate, pre-programmed, stimulus response mechanism available to the individual. Innate intelligence is not, however, defined by degree. Complex, essentially reactive, models have been developed to comprehensively describe and (so largely) explain the behavioral repertoire of several non-primate vertebrate species, including small mammals, birds and fish. This is clearly seen in the pioneering work of ethologists such as Baerends (1976), Lorenz (1950), and Tinburgen (1951). Travers (1989) presents a computer simulation of the stickleback's innate reproductive behavior. Hallam, Hallam and Halperin (1994) a simulation of aspects of behavior in the Siamese fighting fish. Several schemes for partitioning behaviors have been proposed. These include Tinburgen's (1951) *Innate Releaser Mechanism* (IRM) hierarchical scheme and Brooks' (1986) robot *subsumption architecture*. Tyrrell (1993) provides a useful summary of a variety of action selection mechanisms drawn from natural and artificial examples. This mapping of stimulus situation to action response will be referred to here as the *Static Policy Map* (SPM) for an animal or animat.

Behaviorist learning is considered to be *reinforcement*, or strengthening of the activating bond between stimulus and response (e.g. Hull, 1943, for an extensive treatment of experimentally induced reinforcement in animals) in the presence of a *rewarding* outcome – the *law of effect*. That is, the occurrence of a desirable event concurrently (or immediately following) an application of the S-R pair enhances the likelihood that the pairing will be invoked again over other alternative pairings, conversely, with a reduced likelihood for undesirable (*aversive* or *punishing*) events. In principle, new pairings may be established by creating an S-R link between a stimulus and a response that were active concurrently with (or immediately preceding) the rewarding event.

Reinforcement Learning (RL) methods (Kaelbling *et al*., 1996; Sutton and Barto, 1998) propose that reward may be propagated from states that are directly associated with reward to states that are not in order to construct an optimal state-action policy map to achieve reward in the long term. Reinforcement learning techniques are grounded in a heritage long established by the dynamic programming, learning automata and optimization research communities (e.g. Bellman, 1957; Howard, 1960). Many of the formally justifiable results in RL arise from modeling in a *Markov Decision Process* (MDP) environment (Howard, 1960). The Markov condition refers to the

independence of states, each from the others, and from actions taken by the agent. In a variant, the *Partially Observed MDP* (*POMDP*), the agent is unable to reliably determine the current state it is in. *Actor-critic* models (Barto, 1995; Sutton and Barto, 1998) separate the behavior selection function (Actor) from the learning-signal function (Critic), which typically compares the actual reward with an estimate, to then update the policy settings. Barto (1995) proposes RL as a model of learning and action selection in the Basal Ganglia, though this view is not itself without critics (Dayan and Balleine, 2002; Khamassi, Lachèze, Girard, Berthoz and Guillot, 2005). Sutton (1991) has proposed RL as a model of animat maze following and this is considered further in section 9.

2.2 The Associationist Model

A second theoretical position, broadly characterised by the term *associationist* (Bower and Hilgard, 1981, Ch. 2), concerns the direct associability and anticipation of cell clusters or stimuli following repeated pairings of activations. Learning in this model is often referred to as Hebbian Learning (after Hebb, 1949; see also a recent perspective on this work by Sejnowski, 1999). The adaptive properties of correlated activations between neural cell assemblies has been extensively studied as an indicator of underlying neural mechanisms (e.g. Bi and Poo, 2001, for a recent review). Several theoretical models have been proposed based on Hebbian learning principles (Sejnowski, Dayan and Montague, 1995), and been employed in a number of application areas, such as covariance learning (Minai, 1997; Sejnowski, 1977) and principal component analysis (Oja, 1992). While of greater significance in other aspects of animal and brain modelling, this approach does not specifically incorporate an action component and discussion of it will be restricted here to a supporting role in the action selection context.

2.3 Classical Conditioning

A third, deeply influential, approach to animal learning developed during the 1920s as a result of the work of Ivan Pavlov (1849-1936), now usually referred to as *classical conditioning*. The procedure is well known and highly repeatable. It is neatly encapsulated by one of the earliest descriptions

provided by Pavlov (1927). Dogs naturally salivate in response to the smell or taste of meat powder. Salivation is the unconditioned reflex (UR), instigated by the unconditioned stimulus (US), the meat powder. Normally the sound of (for instance) a tone does not cause the animal to salivate. If the tone is sounded almost simultaneously with the presentation of meat powder over a number of trials, it is subsequently found that the sound of the tone alone will cause salivation. The sound has become a conditioned stimulus (CS). The phenomena is widespread, leading Bower and Hilgard (1981, p. 58) to comment "*almost anything that moves, squirts or wiggles could be conditioned if a response from it can be reliably and repeatably evoked by a controllable unconditioned stimulus*". One might conjecture that, if a simple reflex provides protection or preparation for the animal, then anticipation of that need by the CS will pre-empt the outcome to the animal's advantage.

The conditioned response develops with a characteristic sigmoid curve with repeated CS/US pairings. Once established, the CS/UR pairing diminishes if the CS/US pairing is not regularly maintained, the *extinction* process. We may note that the scope of the US may be manipulated over a number of trials to either be highly differentiated to a specific signal, or conversely gradually generalized to respond to a range of similar signals (for instance, a tone of particular frequency, versus a range of frequencies about a center). Higher-order conditioning (Bower and Hilgard, 1981, p. 62) allows a second neutral CS' (say, a light) to be conditioned to an existing CS (the tone), using the standard procedure. CS' then elicits the CR. Long chains are not, however, easy to establish or maintain.

The classical conditioning procedure is highly repeatable and is easily demonstrated across a wide range of reflexes and species. It has been extensively modelled both by implementation and mathematically. Rescorla and Wagner (1972) produced an influential classical conditioning model relating the effects of reinforcement in conditioning to associative strength – the degree to which the CS is "surprising". Sutton and Barto (1987; 1990) extend the Rescorla and Wagner model of classical conditioning using an actor-critic architecture based on the temporal difference RL model. Recently, Courville, Daw and Touretzky (2006) have re-cast classical conditioning into a Bayesian interpretation. See also Balkenius and Morén (1998), and Vogel, Castro and Saavedra (2004) for recent reviews of classical conditioning models.

2.4 Operant Conditioning

A radically alternative view of learning was proposed by B.F. Skinner (1904-1990), that of *instrumental* or *operant conditioning* (Bjork, 1993, for a review of Skinner's life and work). In this model, responses are not "elicited" by impinging sensory conditions, but "emitted" by the animal in anticipation of a reward outcome. Reinforcement strengthening is therefore considered to be between response (R) and rewarding outcome (O), the R-O model, not between sensation and action, as in the S-R model. One might conjecture that the instrumental approach underpinning this effect allows the animal to learn, unlearn and take actions that achieve specific needs on a direct basis.

Operant conditioning is illustrated by reference to an experimental apparatus developed by Skinner to test the paradigm, now universally referred to as the "Skinner box". In a typical Skinner box experiment the subject animal (typically a rat) operates a lever to obtain a reward, say a small food pellet. As a preliminary step, the subject animal must be prepared by the experimenter to associate operating the lever with the food reward. However, once the subject is conditioned in this manner the apparatus may be used to establish various experimental procedures to investigate effects such as stimulus differentiation, experimental extinction, the effects of adverse stimuli ("punishment schedules") and the effects of different schedules of reinforcement (such as varying the frequency of reward). As the apparatus may be set up to automatically record the activities of the subject animal (lever pressing), long and/or complicated schedules are easily established.

Operant conditioning has found application in behavior "shaping" techniques, where an experimenter wishes to directly manipulate the overt behavioral activities of a subject, animal or human. In the simplest case the experimenter waits for the subject to emit the desired behavior and immediately afterwards presents a reward (a rat must be prepared in this way before it can be used in a Skinner box). Importantly, it is to be noted that the R-O activity may be easily manipulated so as to occur only in the presence of a specific stimulus, which may in turn be differentiated or generalized by careful presentation of reward in the required circumstances.

This has lead to the assertion that operant conditioning is properly described by as three-part association, S-R-O. The stimulus (S) itself now appears to act as a *conditioned reinforcer*, where it

previously had no inherent reinforcing properties. In turn, a new response in the context of another stimulus (Sy) and response (Ry) may be conditioned to the existing triple (Sx-Rx-O):

Sy-Ry-Sx-Rx-O

Chains of considerable length and complexity have been generated in this way. They have been used, for instance, in the film industry to prepare performing animals (Bower and Hilgard, 1981, pp. 178-179). It is, of course, a given that the rewarding outcome is itself a sensory event with direct (innate) association with some condition the subject wants (or in the case of aversive condition, does not want). For instance, if the subject animal is not hungry when offered food, the connection will not be manifest and might not be formed. It is also the case that an apparently non-reinforcing sensory condition can attain reinforcing properties if presented in conjunction with an innately reinforcing (positive or negative) one, the *secondary* or *derived reinforcement* effect (Bower and Hilgard, 1981, p. 184). Derived reinforcers will also condition responses unrelated to the original one.

Whilst enormously influential in its time, only a relatively small number of computer models directly follow this approach (e.g. Saksida, Raymond and Touretzky, 1997). Schmajuk (1994) implements Mowrer's (1956) "two-factor" theory, incorporating both classical and operant conditioning effects. Bindra (1972) has also presented a formulation to combine these two approaches. Dayan and Balleine (2002) discuss issues of reward and motivation in a classical and instrumental conditioning based on an RL formulation in a neurological context.

2.5 Sign-learning

The final model to be considered is derived from Tolman's (1932) notion of a *Sign-Gestalt Expectancy*, a three part "basic cognitive unit" of the form S1-R-S2, in which the occurrence of the stimulus S1 in conjunction with the activity R, leads to the expectation or prediction of the outcome S2 (which may or may not be "rewarding"). This is largely equivalent to Catania's (1988) description of the fully discriminated operant connection as a *three-part contingency* of "stimulus – response – consequence", but with the essential difference that it is the identity of the outcome that is to be recorded, rather than just a measure of the desirability or quality of the connection as assumed in

purely S-R behaviorist or operant conditioning approaches. The formulation developed here is derived from a set of 13 postulates originally devised by MacCorquodale and Meehl (1953), later refined by Witkowski (2003). These encapsulate the main tenets of Tolman's theoretical position in a manner tractable to modeling and analysis. The interpretation adopted here emphasizes the predictive nature of the link and this is used both in generating behavior sequences and to drive the learning process.

Tolman was a pioneer of what is now the established *cognitive school* within psychology (see MacCorquodale and Meehl, 1954, for a broader retrospective of Tolman's work). His aims included being able to demonstrate that animals were capable of both explicit problem solving and "insight", that they could anticipate outcomes of their actions and so avoid potential problems. He instigated a number of ingenious experiment types, among them latent learning and place learning (Bower and Hilgard, 1981, ch. 11), that remain key differentiators between the properties of the S-R behaviorist and expectancy components of the Calculus to be presented here. These experiment types are considered later in the paper (section 9) as part of a discussion of the role of Reinforcement Learning (RL, section 2.1) and the anticipatory approach. Tolman's means-ends approach inspired and remains one of the central techniques of Artificial Intelligence problem solving and planning techniques (e.g. Russell and Norvig, 2003, sections II and IV). The Dynamic Expectancy Model (DEM) (Witkowski, 1998; 2000; 2003) and the Anticipatory Classifier System (ACS) model (Stoltzmann, Butz, Hoffmann and Goldberg, 2000) represent recent three-part action selection and learning models. Butz, Sigaud and Gérard (2003) present a summary of recent explicitly anticipatory models.

3 Sense, Action and Valence

For largely historical reasons, sensations are widely referred to as *stimuli* in this body of literature and the actions or behaviors generated as *responses*. This is not entirely satisfactory, as it largely fails to capture the range of interpretations required by the five factor theories taken together. Consequently, this paper will refer to the sense-derived component as a *sensory signature* or *Sign* and denote such events by the symbol S. Sub-scripts will be used to differentiate Signs where necessary. Equally, the term "response" seems pejorative and the more neutral term *Action* will be preferred, similarly abbreviated to A. Each Action will have associated with it an *action cost*, *ac*, indicating the time,

effort or resource required to perform it. Any Action may also be assigned an *activation level*, determined according to the rules presented later. Once activated, an Action becomes a candidate for *expression*, in which the Action is performed by the animal and may be observed or measured directly.

A Sign is defined as a conjunction of detectable conditions (or their negations, acting as inhibitory conditions), typically drawn directly from the senses. Any Sign where all the conditions currently hold is said to be *active*. A Sign may be activated by some very specific set of detected sensory conditions, or be active under a wide range of conditions, corresponding to differentiated or generalized sensing.

Signs and Actions may be taken as representational placeholders for activations in the brain. It has long been known that there is a close mapping between the external afferent (sensory) and efferent (motor) pathways and specific areas of the cerebral cortex. For instance, the sensory and motor "homunculi" of Penfield and Rasmussen (1950), between the eye and highly differentiated responses in the visual cortex (Hubel and Wiesel, 1962), "place" or "grid" representations of locations in physical space in the hippocampus (O'Keefe and Nadel, 1978) and medial entorhinal cortex (Hafting, Fyhn, Molden, Moser and Moser, 2005), or barrel representations of whisker activation in the rat somatosensory neocortex (Welker, 1976; Leergaard, Alloway, Mutic and Bjaalie, 2000). Such mappings appear to be ubiquitous within the mammalian brain (Kaas, 1997) and are no doubt represented to a greater or lesser extent in other vertebrates.

Any Sign that is anticipated, but not active, is termed *sub-active*. Sub-activation is a distinct condition from full activation. It is important to distinguish the two, as the prediction of a Sign event is not equivalent to the actual event and they have different propagation properties.

Additionally, any Sign may assume a level of *valence* (after Tolman, 1932), the extent to which that Sign has goal like properties, indicating that it may give the appearance of driving or motivating the animal to directed action selection behavior. Valence may be positive (goal seeking or rewarding) or negative (initiating avoidance behaviors or being aversive). A greater valence value will be taken as more motivating, or rewarding, than a lesser one. Some Signs will hold valence directly, some via propagation from other Signs holding valence, *sub-valence*.

While long-standing notions of what constitutes motivation in humans (e.g. Maslow, 1987) and animals (e.g. Bindra and Stewart, 1966) remain contentious, recent studies have indicated several areas of the (mammalian) brain are directly implicated in initiating or sustaining overt behavior types (Sewards and Sewards, 2003, for review). They identify areas correlating strongly with primary drives: hunger, thirst, sexual activity, fear, nurturance of infants, motivational aspects of pain, sleep and social dominance, by the strong correlation of stimulation (or disruption to brain tissue by lesion) with activation or suppression of behaviors associated with these "drive" types. These areas may themselves be affected, directly or indirectly, by physiological change (hunger, thirst), hormonal balance (sexual activity) or by connection to other brain areas. It seems likely that these areas, or those mirroring them, also have sensory attributes.

As with activation and sub-activation (anticipatory), the valence and sub-valence (motivating) properties may also be propagated between Signs under the conditions described in section 5. In the SRS/E implementation model, any Sign that is a direct source of valence is deemed *satisfied* once it has become active and its motivating valence properties are automatically cancelled.

4 The Forms of Connection

This paper proposes that the principal effects of the five factor theories can be adequately explained by the Action Selection Calculus by adopting a combination of three connection types and that their underlying function is to provide a temporally predictive link between different Sign and Action components. This section describes the structure and principal properties of the three connection types used. Section 5 describes the way in which prediction, activation and valence propagates across these different connection types. Section 6 then describes the way in which new connections are formed and the strength of existing connections modified by anticipatory learning processes.

While noting that the model described here is highly abstracted, its biologically inspired background grounds it in the notion that, in nature, these abstract links represent physical neural connections between parts of the animal's nervous system and brain. These links, and such properties as sub-activation and valence, represent conjectures (from experimental observation) about the function of the brain that may be corroborated or refuted by further investigation. Note that these

connection type definitions are derived from observation of external behavior, not inference about what function neural structures might serve.

Two of the abstract link types proposed below are bi-directional. Propagation effects across these links are asymmetric and these properties are discussed in section 5. This is not intended to imply that "bi-directional neurons" are necessary, only that the structures that construct these linking elements have a complexity suited to the task. Where the animal does not possess a link or type of link (on the basis of its genetic makeup) it will be congenitally incapable of displaying a corresponding class of action selection behavior or learning. Of course, there are many other possible connection formats between arbitrary combinations of Signs and Actions; but it will be argued that these are sufficient to explain the principal properties of the five factor theories.

Connection type SA: $S_1 \overset{w}{\rightarrow}_{t\pm\tau} (A \wedge S_V)$

Connection type SS: $S_1 \overset{v,c}{\leftrightarrows}_{t\pm\tau} S_2$

Connection type SAS: $(S_1 \wedge A) \overset{v,c}{\leftrightarrows}_{t\pm\tau} S_2$

While **SA** connections have only an implicit anticipatory role, connection types **SS** and **SAS** are both to be interpreted as making explicit anticipatory predictions. The symbols used in this paper are defined and described below. Their meanings are also summarized in Appendix one.

The **SA** connection is a rendition of the standard S-R behaviorist mechanism, with a forward only link from an antecedent sensory condition initiating (or at least predisposing the animal to initiate) the action A, as represented by the link "→". This symbol should definitely not be associated with logical implication, its interpretation is causal not truth preserving. The symbol *t* will indicate temporal delay (with range "±τ"), which may be introduced between the sense and action parts. The (optional) Sign S_V is postulated as a mechanism for reinforcement learning and is not required where learning across the connection (updating *w*) is not observed. The conjunctive connective symbol "∧" should be read as "co-incident with".

In keeping with standard behaviorist modeling, *w* will stand to indicate the strength, or *weight*, of the connection. This weight value will find application in selecting between candidate connections

and in considering reinforcement learning. Traditionally, the strength of the stimulus and a habituation mechanism for the action would also be postulated (Hull, 1943, for a comprehensive discussion of these and related issues). Specifically, the strength or likelihood of the response action will be modulated by the strength of the stimulus Sign.

4.1 Explicitly Anticipatory Connection Types

Connection type **SS** notates a link between two Signs and indicates that Sign S_1 anticipates or predicts the occurrence of Sign S_2 within the specific time range $t \pm \tau$ in the future. This is indicated by the right facing arrow in the link symbol "⇆". The *valence value*, *v*, of S_1 is a function of the current value of the valence value of S_2 and is therefore associated with the left facing part of the link.

Where the value $t \pm \tau$ is near zero, the link is essentially symmetric, S_1 predicts S_2 as much as S_2 predicts S_1. This is the classical Hebbian formulation (Hebb, 1949; Bi and Poo, 2001). Where *t* is greater than zero (negative times have no interpretation in this context), the link is considered asymmetric and predictive. The assertion that S_1 predicts S_2 is no indicator that S_2 also predicts S_1. As the relationship between the two Signs is not necessarily causal, the animal may hold both hypotheses simultaneously and independently, as separate **SS** connections.

The **SAS** connection differs from **SS** by the addition of an instrumental Action on the left hand side. The prediction of S_2 is now contingent on the simultaneous activation of both S_1 and the action A. The interpretation of the corroboration value *c* and the temporal offset *t* and range τ remain the same. Transfer of valence *v* to S_1 needs to now be a function of both S_2 and the action cost (*ac*) of A. This connection can be read as "the Sign S_2 is anticipated at time *t* in the future as a consequence of performing the action A in the context of S_1". Equally, it may serve as an instrumental operator: "to achieve S_2 at time *x* in the future, achieve S_1 at time *x*-*t* and perform action A". Such links also take the form of independent hypotheses, giving rise to specific predictions that may be corroborated.

The *corroboration value*, *c*, associated with each **SS** or **SAS** link records the probability that the left hand (condition) side of the link correctly predicts the Sign (consequence) on the right hand side. A discounting strategy is used in SRS/E to update the corroboration value gives greater weight to the outcome of recent predictions (whether they succeed or fail) and successively discounts the

contribution of previous predictions. The rate at which this discounting occurs is controlled by two numeric factors α, the *corroboration rate*, and β, the *extinction rate*, respectively controlling the discounting rate for successful and unsuccessful predictions. The generic corroboration value update rule used in SRS/E, incorporating α and β, will be considered in section 6.1.

5 The Forms of Propagation

The five "rules of propagation" presented in this section encapsulate the operations on the three connection types with regard to the five factor theories. The rules define (i) when an Action becomes a candidate for expression, (ii) when a Sign will become sub-activated, (iii) when a prediction will be made, and (iv), when a Sign will become valenced by propagation.

In the semi-formal notation adopted below *active*(), *sub_active*(), *expressed*(), *valenced*() and *sub_valenced*() may be treated as predicate tests on the appropriate property of the Sign or Action. Thus, *active*(S_1) will be asserted if the Sign denoted by S_1 is active. The disjunction "$\vee$" should be read conventionally as either or both, the conjunction "$\wedge$"as "co-incident with". On the right hand side of the rule, *activate*(), *sub_activate*(), *predict*(), *sub_valence*() and *set_valence*() should be taken as "internal actions", operations taken to change the state or status of the item(s) indicated.

Rule P1 Direct Activation:

For each **SA** link,
if (*active*(S_1) $\vee$ *sub_active*(S_1))
then *activate*(A, *w*) $\vee$ *set_valence*(S_x, *v*)

Rule P2 Sign Anticipation:

For each **SS** link,
if (*active*(S_1) $\vee$ *sub_active*(S_1))
then *sub_activate*(S_2)

Rule P3 Prediction:

For each **SS** link,
if(*active*(S_1))
then *predict*(S_2, $t\pm\tau$)

For each **SAS** link,
if(*active*(S_1) $\wedge$ *expressed*(A))
then *predict*(S_2, $t\pm\tau$)

Rule P4 Valence transfer:

For each **SS** link,
if(*valenced*(S_2) $\vee$ *sub_valenced*(S_2))

then *sub_valence*(S_1, *f*(*v*(S_2), *c*))

For each **SAS** link,
if(*valenced*(S_2) ∨ *sub_valenced*(S_2))
then *sub_valence*(S_1, *f*(*v*(S_2), *c*, *ac*(A)))

Rule P5 Valenced activation:

For each **SAS** link,
if(*active*(S_1) ∧ *sub_valenced*(S_1))
then *activate*(A, *v'*)

Rule **P1** (Direct Activation) expresses the standard S-R behaviorist rule. It is included here on the strength of that basis alone, evidential support for this form of link is widespread and not contentious. Only in the simplest of animals would the activation of action A lead to the direct overt expression of the action or activity. As there is no assumption that Signs are mutually exclusive (the Markov property), many actions may become candidates for expression. The simplest strategy involves selecting a "winner" based on the weightings and putting that action forward to the effector system for external expression. In the SRS/E implementation model, goal setting is considered a form of direct action behavior, so the rule is shown permitting the activation of an Action, or the setting of a valence value, *v*, to any nominated Sign, S_x.

Rule **P2** (Sign Anticipation) allows for the propagation of anticipatory sub-activation. The effect is instantaneous, notifying and allowing the animal to modify its action selection strategy immediately in anticipation of a possible future event, such as initiating a conditioned reflex (section 2.2). Evidence for the sign anticipation rule is derived from primary and second order classical conditioning studies, where short chains of apparently anticipatory Sign-stimuli can be established experimentally (section 2.2). Evidence from these studies would further indicate that sub-activation propagates poorly (i.e. is heavily discounted).

Rule **P3** (Prediction) allows for a specific prediction of a future event to be recorded. This calls for a limited form of memory of possible future events, analogous to the more conventional notion of a "memory" of past events. Under this formulation, predictions are created as a result of full activation of the Sign and actual expression of the Action, and are therefore non-propagating. Predictions are made in response to direct sense and action and are employed in the corroboration process (section 6.1). This process is distinct from sub-activation, which is propagating, but non-corroborating. Rule

P3 is a conjecture based on the notion of latent learning (Thistlethwaite, 1951). The act of predicting a future Sign event and its subsequent corroboration providing a source of internal "reward" that is independent of an external reinforcing signal.

Rule **P4** (Valence Transfer) indicates the spread of valence backwards along chains of anticipatory links. The *sub_valence*() process is shown in different forms for the **SS** and **SAS** links, reflecting the discounting process mentioned earlier. As an exemplar, in the SRS/E model valence is transferred from S_2 to S_1 across the **SAS** link according to the generic formulation: $v(S_1) := v(S_2) * (c / ac(A))$. By learning rules **L2** and **L3** (section 6.1) link corroboration, $0 < c < 1$ and action cost $ac(A) \geq 1.0$ (by definition in SRS/E), the valence value associated with the left hand Sign S_1 will always be less than that for the right hand Sign S_2, i.e. $v(S_1) < v(S_2)$. Valence propagates preferentially across high confidence links with "easy" (i.e. lower cost value) Actions. This propagated value may be interpreted as a *cost estimate* of performing the Action. Given that the propagated values relate to discounted predictions, the relationship of this method to Bayesian inference should be noted (e.g. Bishop, 1995).

Valence therefore spreads throughout the network of links starting from any primary source of valence, any S_1 Sign adopting the highest (best) valence value if there are multiple paths to it. Note here that the valence value *v*' refers to the valence value of the Sign holding direct valence (the *top-goal*). This transfer mechanism, implemented as a simple variant of the established A* graph traversal algorithm (Hart, Nilsson and Raphael, 1968; Nilsson, 1971), is straightforward and has proved robust in operation in SRS/E. Valence transfer is rapid (i.e. at propagation rates) and is independent of the predictive timing component $t \pm \tau$. As the sources of valence change, the functional, though not physical, structure of this graph changes also.

This process builds a *Dynamic Policy Map* (DPM), assigning a valence value to each Sign that can be reached from the original source of valence by repeated application of **P4**. The mapping indicates the Action associated with the Sign on the path of minimized total estimated cost to the directly valenced Sign (from each **SAS** link).

Evidence for valence transfer is derived from several sources in the factor theories. Most directly in the conditioned reinforcer effect (section 2.4), in which a reinforcing effect is propagated along a

chain of stimulus-response-stimulus (modeled as **SAS** links) events in animal shaping techniques. It is also a central tenet of the Sign-learning approach (section 2.5), in which valence or motivation is propagated backwards ("means-ends") until a suitable condition is encountered to trigger the activation of an Action (by **P5**). That chains of considerable length may be formed is evidence that valence propagates well under these conditions. The secondary or derived reinforcing effect (section 2.4) provides strong evidence for valence transfer across **SS** links. That valence transfer is maintained by a primary source of valence may be trivially demonstrated by removing the original source of valence and noting that the expression of directed behavior is suppressed.

Rule **P5** (Valenced Activation) indicates the activation of any Action A where the antecedent Sign S_1 is both active and valenced within the current Dynamic Policy Map. As with rule **P1**, many Actions may be affected. The one associated with the highest overall S_1 valence value is selected. The rule may be inferred from the effect of placing an animal at given points (identified by stimulus Signs) in a shaped behavior chain and noting that it selects the Action appropriate to that point in the chain (though care must be taken as there is no guarantee the Markov property, independence of states, holds in animal perception). It is central to the notion of action selection from a dynamic policy map.

The choice process by which the various activated Actions give rise to the action to be selected for overt expression is the subject of section 8. For a simple S-R only (rule **P1**) system, this might be summarized as selecting the action associated with the highest weight value, but there must be a balance between the actions activated by rule **P1** and those by **P5**.

6 The Forms of Learning

This section describes the conditions under which learning will take place. In the action selection model presented, the net effect of learning is to modify the Actions or activities to be expressed (and so the observable behavior of the animal) in response to a particular motivating Sign. Each of the five factor theories takes a particular position on the nature of learning.

In the first, *reward based learning*, learning is taken to be a consequence of the animal encountering a valenced situation following an action – one that is characterized as advantageous/disadvantageous and thus interpreted as "rewarding" (or not) to the animal. This is

frequently referred to as reinforcement learning (section 2.1). There are a wide range of reinforcement learning methods, so a generic approach will be adopted here.

In the second, *anticipatory learning*, "reward" is derived from the success or otherwise of the individual predictions made by the propagation rules given in section 5. In one sense, the use of link type **SAS**, as described here, can be seen as subsuming link types **SA** and **SS**, but the converse does not hold. In the **SA** link, the role of anticipation in the learning process is implicit but is made explicit in the **SS** and **SAS** type links.

Learning rule L1 (the reinforcement rule):

For each **SA** link
if (*active*(A) $\wedge$ (*valenced*(S_V) $\vee$ *sub_valenced*(S_V)))
then *update*(*w*, α)

This is a generic form of the standard reinforcement rule. If the action is followed by any Sign (S_V) that provides valence, then the connection weight *w* will be updated by some proportional factor α. Several well established weight update strategies are available, such as Watkins' *Q-learning* (Watkins and Dayan, 1992) and Sutton's *temporal differences* (TD) method (Sutton, 1988), see Sutton and Barto (1998) for review, section 2.1. In each the net effect is to increase or decrease the likelihood that the link in question will be selected for expression in the future.

6.1 Methods of Anticipatory Learning

A central tenet of the anticipatory stance described in this paper is that certain connective links in the model make explicit predictions when activated. Recall that propagation rule **P3** creates explicit predictions about specific detectable events that are anticipated to occur in the future, within a specific range of times (denoted by $t\pm\tau$). The ability to form predictions has a profound impact on the animal's choice for learning strategies. This section considers the role played by the ability to make those predictions.

Learning rule L2 (anticipatory corroboration):

For each (**SS** $\vee$ **SAS**) link
if(*predicted*(S_2, $-t\pm\tau$) $\wedge$ *active*(S_2))
then *update*(*c*, α)

Learning rule L3 (anticipatory extinction):

For each (**SS** $\vee$ **SAS**) link,
if(*predicted*(S_2, -*t*$\pm\tau$) $\wedge$ $\neg$*active*(S_2))
then *update*(*c*, β)

Learning rule L4 (anticipatory link formation):

For each (**SS** $\vee$ **SAS**) link,
if($\neg$*predicted*(S_x) $\wedge$ *active*(S_x))
then *create_SAS_link*(S_y, A_y, S_x, *t*, τ)
or *create_SS_link*(S_y, S_x, *t*, τ)

These three rules encapsulate the principles of anticipatory learning and are applicable to both **SS** and **SAS** link types. Three conditions are significant. First, where a Sign has been predicted to occur and the predicted Sign occurs at the expected time. The link is considered corroborated and is strengthened (corroboration rule **L2**). Second, where a Sign prediction is made, but the event does not occur, the link is considered dis-corroborated and weakened (extinction learning rule **L3**). Third, where a Sign occurs, but was not predicted. The immediately preceding circumstances are used to construct a new link that would predict the event were the recorded circumstances (S_y, A_y) to reoccur in the future (anticipatory link formation rule **L4**).

The SRS/E computer implementation employs a simple but robust, effective and ubiquitous update function for the link corroboration value, *c*, for each **SS** and **SAS** rule. For each successful prediction made by the link the anticipatory corroboration rule **L2** is invoked. The new value of *c* for the link is given by $c := c+\alpha\ (1-c)$, where $(0 \leq \alpha \leq 1)$. For each failed prediction the extinction rule **L3** is invoked, the new link corroboration value is given by $c := c-\beta(c)$, $(0 \leq \beta \leq 1)$. These update functions are asymptotic towards 1.0 (maximum) and zero (minimum) respectively for successful and failed prediction sequences. The net effect of these update rules is to maintain a form of "running average" more strongly reflecting recent predictions, with older predictions becoming successively discounted, tending to zero contribution. Where no prediction was made by a link, the corroboration value *c* remains unchanged regardless of the occurrence of S_2. This is consistent with the notion that a link is only responsible for predicting an event under the exact conditions it defines.

Particular settings of α and β values are specific to the individual animal. The greater the values of α and β, the more aggressively recent events are tracked and the contribution of past events discounted. The values of α and β may be set arbitrarily for any individual. There are no obvious "optimal" values, the animal appearing, to an observer, more or less persistent in its expressed behavior under different conditions. In the exemplar experimental procedure of section 9.3, the value of β sets the extinction rate for *c* of the failed **SAS** link and so determines the time to action selection policy change. The emulated experiments described in section 9 use (empirically determined) system default values of 0.5 and 0.2 for α and β respectively.

Learning rules **L2** and **L3** reflect the conventional notions of strengthening by reinforcement (**L2**, α) and weakening by extinction (**L3**, β), which is common to each of the forms of learning considered here. The discounting form of the generic rule for SRS/E is ubiquitous throughout the natural (e.g. Bower and Hilgard, 1981) and machine learning (e.g. Kaelbling *et al*., 1996) literature, and in artificial neural networks (e.g. Bishop, 1995).

Where a Sign event occurs, but is currently unpredicted by any link, this is taken as a cue to establish a new link, using the anticipatory link formation rule **L4**. The link is formed between the unpredicted event (as S_2) and some recently active event (as S_1) at time *t*. Where an **SRS** link is created, some expressed Action A_y contemporary with the new S_1 is also implicated. Without any *a-priori* indication as to which new links might be effective, higher learning rates can be achieved by forming many links and then allowing the corroborative learning rules **L2** and **L3** to separate the effective from the ineffective with each successive prediction – competition by corroboration. The choice of how many new links are formed and the range of values for *t* and τ are specific to the individual animal. The SRS/E model incorporates a learning probability rate parameter, λ, which determines the probability of a link being formed given the opportunity to do so (section 9). The learning probability rate reflects the observation that tasks typically take many trials to learn (such as the preliminary step of establishing operant behaviors, section 2.4, prior to operant testing).

Learning by **L4** may proceed from *tabula rasa* and is rapid while much is novel. In a restricted environment, link learning will slow as more is correctly predicted, but will resume if circumstances

change. Link learning in neural tissue does not necessarily imply the growth of new neural pathways, rather the adoption of existing uncommitted ones for a specific purpose. The natural structure of the brain may therefore strongly determine what can and cannot be learned. Shettleworth (1975), for instance, reports that certain behaviors are much more amenable to (classical) conditioning than others in the golden hamster.

No rule for link removal is considered here, but has been discussed elsewhere in the context of the Dynamic Expectancy Model (DEM). Witkowski (2000) considers the rationale for retaining links even when their corroboration values fall to very low values, based on evidence from behavioral extinction experiments (Blackman, 1974). The behavioral extinction properties for operant and instrumental conditioning procedures are substantially different. Extinction in Classical conditioning is typically rapid with successive uncorroborated trials. Instrumental extinction typically occurs only after an extended number of uncorroborated trials (Blackman, 1974). This may be taken as further evidence that the detailed properties of **SS** and **SAS** links are inherently different. Evidence would suggest that extinguished links are not lost, as the conditioning effect can spontaneously reappear after a refractory period.

7 Explaining the Five Factors

This section returns to the action selection factor theories outlined in section 2 (associationism is given no further treatment here, except insofar as it supports the others). Each is discussed in turn in terms of the link types, propagation rules and learning rules presented and discussed in sections 4, 5 and 6. As previously indicated, each theory supports and is supported by an (often substantial) body of experimental evidence, but that each theory in turn fails to capture and explain the overall range of action selection behaviors displayed by any particular animal or species. The conceptually simpler approaches are covered by single links and rules; some require a combination of forms, yet others are to be re-interpreted in the light of this formulation.

7.1 Stimulus-Response Behaviorism and Static Policy Maps

With no embellishments, S-R behaviorism is reduced to connection type **SA** and propagation type **P1**. The underlying assumption of these strategies is to tailor the behavior of the organism, such that the

actions at one point sufficiently change the organism or its environment such that the next stage in any complex sequence of actions becomes enabled and is indicated by the ensuing sensory conditions. This is the *static policy map*. SRS/E records these connections in a list, effectively ordered by the weight parameter, *w*. Recall that the weighting value *w*, and hence the ordering, may be modified by reinforcement learning (Sutton, 1988; Sutton and Barto, 1998; Watkins and Dayan, 1992; section 2.1). Static policy maps should not, therefore, be thought of as unchanging or unchangeable. Some behaviors appear completely immune to modification by learning, such as the egg recovery behavior of the greylag goose (Tinbergen, 1951), others modifiable by learning to varying extents (Shettleworth, 1975) according to function and apparent purpose.

Given a sufficient set of these reactive behaviors, the overall effect can be to generate exceptionally robust behavioral strategies, apparently goal seeking, in that the actually independent elements of sense, action and actual outcome combinations, inexorably leads to food, or water, or shelter, or a mate (Bryson, 2000; Lorenz, 1950; Maes, 1991; Tinbergen, 1951; Tyrrell, 1993). The key issue here is that the animal need have no internal representation of the potential outcome of the action it takes (Brooks, 1991b); the circumstances suited to the next stage in the chain of behavior arising purely as a consequence of the previous action. If the chain is broken, because the action fails to lead to the next appropriate situation, other elements of the policy will be expressed. In some cases, such as in the courtship rituals of some avian species, a clearly identifiable set of *displacement behaviors*, such as preening or aggression, may be noted when this occurs (Tinbergen, 1951).

Such strategies can appear remarkably persistent and when unsuccessful, persistently inept. Any apparent anticipatory ability in a fixed S-R strategy is not on the part of the individual, but rather a property of the species as a whole. With sufficient natural diversity in this group strategy, these strategies can be very robust against moderate changes in the environment, at the expense of any individuals not suited to the changed conditions.

7.2 Classical Conditioning

Reactive behaviorism relies only on the direct activity of the Sign S_1 to activate A, this is the *unconditioned stimulus* (US) to the *unconditioned response* (UR): the innate reflex. As reflexes are

typically unconditionally expressed (i.e. are localized or have high values of *w*) the US invariably evokes the UR. Rule **P1** allows for sub-activation of the S_1 Sign. Therefore, if an anticipatory **SS** connection is established between a Sign, say S_X and the US Sign S_1, then activation of S_X will sub-activate S_1 and in turn evoke A, the *conditioned response* (CR).

Note the anticipatory nature of the CS/US pairing (Barto and Sutton, 1982), where the CS must precede the US by a short delay (typically <1s). The degree to which the CS will evoke CR depends on the history of anticipatory pairings of S_X and S_1. It is dynamic according to that history, by the corroborative learning rules **L2** and **L3**, the rates depending on the values of α and β. If the link between CS and US is to be created dynamically, then learning rule **L4** is invoked. The *higher order conditioning* procedure allows a second neutral Sign (S_Y) to be conditioned to the existing CS (S_X), using the standard procedure: S_Y now evokes the CR. This is as indicated by the propagation of sub-activation in **P2**.

Overall, the classical condition reflex has little impact on the functioning of the policy map of which its reflex is a part. Indeed, the conditioned reflex, while widespread and undeniable, could be thought of as something of a curiosity in learning terms (B.F. Skinner reportedly held this view). However, it provides direct, if not unequivocal, evidence for several of the rule types presented in this paper.

7.3 Operant Conditioning

Operant conditioning shapes the overt behavior of an animal by pairing the actions it takes to the delivery of reward. The experimenter need only wait for the desired action and then present the reward directly. This is typified by the *Skinner box* apparatus, in which the subject animal (typically a hungry rat) is trained to press a lever to obtain delivery of a food pellet reward (section 2.4). This link is interpreted as an anticipatory one. The action anticipates the sensory condition (food), which, as the rat is hungry, holds valence. Further, the experimenter might present the food only when the action is taken in some circumstances, not others. The animal's behavior becomes *shaped* to those particular circumstances. These are the conditions for the **SAS** connection type. This is equivalent to Catania's (1988) notion of an operant *three-part contingency* of "stimulus – response – consequence".

The association between lever (S_1), pressing (A) and food (S_2) is established as a **SAS** link by **L4**. When the action is preformed in anticipation of S_2, the link is maintained, or not, by **L2** and **L3** according to the outcome of the prediction made (**P3**). While food (S_2) retains valence and the rat is at the lever, the rat will press the lever (**P5**). In the absence of any alternative it will continue to do so. Action selection is now firmly contingent on both encountered Sign and prevailing valence.

Due to valence transfer rule (**P4**) such contingencies propagate. Were the rat to be in the box, but not at the lever, and some movement A_M would take to rat from its current location S_C to the lever S_L, then the **SAS** contingency $(S_C \wedge A_M) \leftrightarrows S_L$ would propagate valence to S_C from S_L and result in A_M being activated for expression. Once the rat is satiated, the propagation of valence collapses and the expression of these behaviors will cease.

The valence propagation rule **P4** allows for *secondary* or *derived reinforcement* effects (Bower and Hilgard, 1981, p.184), in which a normally non-reinforcing Sign may be paired with an innately valenced one across an **SS** link. The valence propagation rule **P4** is also consistent with the *secondary* or *derived reinforcement* effect (section 2.4). This allows for the establishment of behavior chains over successive **SAS** links, where the sequence is built backwards one step at a time from the primary source of valence.

7.4 Tolman's Sign-learning Model

Where the Skinner box investigates the properties of the individual **SAS** link, which may be explored in detail under a variety of different schedules, Tolman's work primarily used mazes. Rats, in particular, learn mazes easily, recognize locations readily and are soon motivated to run mazes to food or water when hungry or thirsty. Mazes are also convenient experimentally, as they may be created with any desired pattern or complexity.

Choice points and other place locations (section 3) in the maze may be represented as Signs (a rat may only be in one location at once, though this may also be incorrectly or ambiguously detected) and traversal between them as identifiable Actions. Every location-move-location transition may be represented as an anticipatory **SAS** connection. Recall that these links are only hypotheses – errors, or imposed changes to the maze are accommodated by the learning rules **L2**, **L3** and **L4**.

It is now easy to see that, when placed in a maze, the animal may learn the structure as a number of **SAS** connections with or without (i.e. latently, section 9) the presence of valence or reward. Novel locations encountered during exploration invoke **L4** and the confidence value *c* is updated each time a location is revisited, by corroboration learning rules **L2** or **L3**. Once encountered, food, when the rat is hungry may impart valence to a location (Link **SS**, by **P4**).

7.4.1 Dynamic Policy Maps

If at any time a location becomes directly or indirectly linked to a source of valence (i.e. food to a hungry rat), this valence will propagate across all the **SAS** (and indeed **SS**) links to establish a *Dynamic Policy Map* (DPM). This takes the form of a directed graph of all reachable Signs. In SRS/E this is considered as a form of modified breadth first search (section 5), in which each Sign node is assigned the highest propagated valence value. Again, this generic "spreading activation" process, as implemented in SRS/E, is both computationally fast and robust in operation. The dynamic policy map is distinguished from the static map by virtue of being constructed on demand from independent links. An identical set of links may give rise to a completely different action selection policy according to the prevailing source of valence.

Once created, each Sign implicated in the DPM is associated with a single Action from the appropriate **SAS** link, the one on the highest value valence path, indicating its current rank in the dynamic policy map. Given this one to one ordered mapping an action may be selected from the DPM in a manner exactly analogous to a static policy map. In this respect, the behavior chaining technique described in section 2.4 looks to be no more than an attempt to manipulate the naturally constructed dynamic policy to favor one sequence of actions to all the others.

The dynamic policy map must be re-evaluated each time there is a change in valence or any learning event takes place (i.e. almost everytime). Sometimes this has little effect on the observable behavior, but sometimes has a dramatic and immediate effect, with the animal reversing its path or adopting some completely new activity, figure 4, section 9, illustrates this.

8 Combining Static and Dynamic Policy Maps

For any animal that displays all the forms of action selection, it becomes essential to integrate the effects of innate behaviour, the static policy map, with the valence driven dynamic policy map. The dynamic policy map is transient, created only when a Sign has valence (goal like properties) and must interleave with the largely permanent static policy map. This may be achieved by postulating a single *subsumption point* (after Brooks, 1986) switching candidate actions for expression between the static and dynamic policy maps.

The numerical valence value of the original valence source (*v*' from section 6, the top-goal) is equated to the range of numerical **SA** connection weight values, *w*. While this numerical value of the top-goal *v'* is greater than any static action ($v' \geq active(w)$), actions are selected only from the DPM and static map selection is suppressed. If at any point $v' < active(w)$, say because a higher priority condition has arisen, DPM action selection is suspended and actions are again taken directly from the static policy. Several papers in Bryson, Prescott and Seth (2005) indicate the Basal Ganglia as the neural seat of this "winner-take-all" selection policy. SRS/E postulates that there are always low-priority exploratory (random) actions to be taken when there are no other candidates for static or dynamic maps; inactivity or rest may also be considered as viable candidates.

This allows for high-priority innate activities, such as predator avoidance, to invariably take precedence over goal directed activities. As the valence of the goal task increases, say with increasing hunger, the chance of it being interrupted in this way decreases. After an interruption from static policy map actions, valenced action selection from the dynamic policy map resumes. The DPM must be reconstructed, as the animal's situation will have been changed and static policy actions may also have given rise to learned changes.

As behavior patterns become more complex, the notion of a simple ordered static policy map model becomes increasingly problematic. Groups of closely related actions need to be coordinated. A case in point would be a typical vertebrate breeding sequence (Lorenz, 1950; Tinbergen, 1951), of say, territory acquisition, courtship, mating, nest building and rearing; each stage being intricate within itself, leading to the next if successful, but also being completely inappropriate at any other

time. The completion of one stage then leaves the animal in a circumstance suitable for the activation Sign of the next stage, and so on. Interleaved with these activities, the animal must still feed, sleep and defend itself from predators or competitors.

Tinbergen (1951) proposed the use of hierarchical *Innate Releaser Mechanisms* (IRM) to achieve this. In each case, the releasing enabler should take its place in the static ranking precedence, with all its subsidiary SR connections simultaneously enabled, but then individually ranked within that grouping for final activation by stimulus.

Maes (1991) proposes an alternative approach (later refined by Tyrrell, 1994) in which behavior strategies are defined in terms of innate, unlearnt three part (**SAS**) links joining sources of motivation to an ordered chain of *appetitive* (seeking) and *consummatory* (achieving) action types. Unlike a dynamic policy map, such mappings are fixed but become enabled *en-bloc* by valencing a single motivating Sign, by the repeated application of valence propagation rule **P4**. This represents a plausible evolutionary intermediate step between pure S-R action selection and learned, dynamically activated, policy maps.

9 Evaluating Static and Dynamic Policy Maps in the Action Selection Calculus Context

To illustrate some of the issues that arise from the Action Selection Calculus this section will look at the role of reinforcement and action selection that arises from the use of static and dynamic policy maps. The example chosen will be a discussion of a simple maze learning task, as might be performed by a rat in the laboratory.

The question under consideration is the role of external reinforcement and the distinction in behavioral action selection terms between reinforcement learning strategies and the dynamic policy map strategy. Three illustrative experiments will be discussed. First, the latent learning procedure, which is used to investigate the role of explicit reward in learning. This is considered to be a classic test to determine whether a task is being performed by an animal within a static (RL) or dynamic policy regime. Second, this is complemented by a discussion of the potential effect of external reward in the context of the anticipatory learning. The third describes a “portmanteau” procedure illustrating

several aspects of valenced operation and DPM construction, combining a number of different existing procedures.

The basic form of the simulation follows that defined by Sutton (1991). The animat is placed in a grid based maze area, in which the animat may make one move Action at each cycle, Up, Down, Left or Right to an adjacent cell. Various cells may be blocked off and the animat will not move if it attempts to enter one of these cells, or if it encounters the boundary. Each grid square is represented by a location Sign, this is consistent both with the requirements of the MDP representation (section 2.1) and for SRS/E (appendix two).

9.1 Latent Learning

Figure 1 replicates the results of a classic latent learning experiment (Witkowski, 1998; after Tolman and Honzig, 1930), indicating that external reinforcement is not required for learning (section 2.1). Tolman argued that if reward were *required* for learning, then a hungry rat that was allowed to explore a maze would have no cause to learn the structure of the maze if there was no reinforcement. This is the situation with, say, *Q-Learning* (Watkins and Dayan, 1992; section 2.1), where a source of reward is mandatory. Reinforcement Learning is predicated on the notion that actions are selected on the basis of predictions of reward. Learning cannot proceed in the absence of any reward injected into the system[4]. Learning by anticipation (rules **L2** and **L4**) is independent of reward, predicated only on the success and failure of internal predictions. The structure of the maze that is learnt, "reward" (food) only providing the motivating valence and expression (propagation rules **P4** and **P5**) once it is encountered.

In the original experiment, three groups of hungry rats (simulated here as valenced animats) are made to run a maze from a start point (bottom) to a constant end location (top) 20 times, (one trial per day in the original rat experiment). For group one, the end location (top) always contains food. For group two, the end location does not contain food until the eleventh trial. For the control group three, the end point never has food.

*** FIGURE 1 ABOUT HERE *** *** FIGURE 2 ABOUT HERE ***

The differential prediction for this experiment is clear. If the animal is employing reinforcement learning, a gradual improvement in performance is expected by group one from the first trial; similarly a gradual improvement by group two, starting at day 11, but only once reward is introduced. The control group is not expected to show any significant improvement over the complete experiment. If, on the other hand, anticipatory or latent learning has occurred, group two will show a clear and immediate improvement in performance once food is encountered and a DPM can be constructed.

Note that in figure 1 it takes on average about 400 random steps to traverse the maze, 11 steps for the shortest route (log scale). Figure 1 is from the animat emulation, but exactly mirrors the changes in performance noted in the original experiments (traces are averages of 100 experimental runs, each with a different random starting seed). The simulation learning parameters are chosen to approximate the animal learning rates ($\lambda = 0.25$). Group one show gradual improvement as they learn the maze, as expected from a reinforcement-based interpretation (and consistently with anticipatory learning). Group three, the control, continue to wander the maze, showing no particular change throughout. Group two, however, model the control group until the twelfth day, but then show a marked improvement in performance once the reward has been discovered. It is interesting to note that the performance of group two exceeds that of group one on the 12^{th} trial. Individuals in group one are obliged to pursue the food by the best route they have at each trial, limiting exploration of better alternatives, but group two are able to explore the maze thoroughly – the *explore-exploit* tradeoff (Kaelbling *et al.*, 1996).

9.2 The Role of Reward in Anticipatory Learning

Latent learning experiments demonstrate that learning in rats need not *depend* on external reinforcement, at least under these conditions, and that this can be emulated in animats also using the rules of the Action Selection Calculus. However, anticipatory learning need not be *independent* of reinforcement by reward. Figure 3 shows the speed-up effect of biasing link learning (*Valence Level Pre-Bias*, VLPB) by **L4** in a similar (simulated) animat maze learning task as group two (section 9.1), such that every opportunity is taken to learn a link (link formation rule **L4**) if it is or has ever been associated with valence (learning probability rate $\lambda = 1.0$), but only occasionally (10%, $\lambda = 0.1$)

otherwise. In the without VLPB trace $\lambda = 0.1$ unconditionally. Learning is significantly more rapid under the VLPB conditions. It makes much sense to bias learning resources to those things immediately relevant to motivated behavior and this postulated mechanism allows an organism to balance speculative latent learning with task related learning. The traces shown in Figure 3 are averages of 100 trial runs, each with a different starting seed. All control trials were averaged to obtain a single representative value.

*** FIGURE 3 ABOUT HERE ***

9.3 Effects of Valence on Action Selection Behavior

This procedure illustrates a number of direct consequences of applying the Action Selction Calculus as described here to an animat under differing valence levels and changing environmental conditions. The first stage emulates the essential aspects of a classic "T-Maze" experiment (e.g. Dayan and Balleine, 2002), in which the path choice selected (or branch of the 'T' in the original) is determined solely and immediately by changing needs. The second stage illustrates the effect of the extinction rule (**L3**) to react to environmental change in a timely manner.

In this set-up the maze (figure 4a) has a start location (center bottom) and two potential sources of valence G_F (food, say), top right and G_W (water), top left. The animat, now fully fed and watered, is allowed to explore the maze shown (after Sutton, 1991) for 1000 action cycles. As neither G_F nor G_W have valencing properties (the animat is neither hungry nor thirsty) no Dynamic Policy Map (DPM) is formed and Actions are selected at random for exploration. This is a latent learning phase (section 9.1). Imagine now the animat is made "hungry" (i.e. valence is associated with G_F, set manually here, by the experimenter) and returned to the start point. A DPM is formed (figure 4b) and the policy path is clearly indicated to location G_F via the shorter route B. Now imagine the animat is fed, but made thirsty (i.e. valence is associated with G_W) and returned to the start point. The new DPM formed indicates the action path directly to G_W, via the shorter path A (figure 4c). The choice point decision is totally different but no new learning has taken place between these two trials.

Now consider a situation where the animat is hungry but not thirsty (i.e. G_F has valence), but location B is now blocked. The recomputed DPM will still indicate a path via B (the blockage is undiscovered), figure 4d. As the intended (up) action to B now fails over several attempts, the anticipatory link to B is extinguished (rule **L3**, at a rate determined by β). DPM alters to advantage the longer path via place A, this is apparent in the simulation as the "preferred action" indicator arrows at each maze place are observed to change direction towards the new preferred path. When this reaches the point of obstruction (after 15 extinction Actions), the observable behavior of the animat abruptly changes as the DPM is reconstructed and a new longer path via A is taken, 4e. This rapid change of policy is not predicted by RL. A RL policy map is constructed iteratively and will adapt only slowly to this block in the path (but see Sutton, 1991, for manipulation to the explore-exploit tradeoff to somewhat improve this).

*** FIGURE 4 ABOUT HERE ***

10 Discussion

This paper has proposed the notion of different types of policy map operating within the animal, static and dynamic and discussed how they may be combined to exhibit apparently different behavioral phenomena under the variety of circumstances the animal may encounter in nature or the laboratory. The Dynamic Expectancy Model has been employed as an implemented (SRS/E) framework for this discussion and used to perform some illustrative experiments.

Perhaps one of the enduring controversies surrounding behavior selection strategies, why it has been so difficult to decide definitively whether actions are selected solely on the basis of sensory input or as a combination of motivation and sensory input, is partially resolved by the comparison of static to dynamic policy maps. Both are predicated on the notion that actions are selected on the basis of sensory input. Key differentiators include whether the same sense modality always gives rise to the same action, or varies according to the current and varying needs of the organism, and how learning is achieved. It is easy to interpret either situation as "motivated". In one case, the latter, this is so. In the static case, the sequence of actions taken are easily interpreted as being motivated by an observer, as they appear to follow a course of actions leading directly to a physiologically necessary outcome. The

dynamic case is clearly demonstrated in maze like tasks where one apparently identical decision point calls for different actions according to the level of motivation (left for thirst, right for food, for instance), section 9.3.

Differences in strategy conventionally hinge on the necessity of external reinforcing reward during learning. Learning in S-R behaviorism depends on the presence of reward, which may (or may not) be propagated to update otherwise static policy maps. External reinforcement of this form is, of course, not required for anticipatory learning. Classic latent learning experiments (section 9.1) provide clear indication that external reward is indeed not required for learning. The valence level pre-bias experiment reminds us that anticipatory learning need not be independent of reward (section 9.2).

Section 8 identified three possible stages in the "evolution" of the policy map. Initially where each reactive **SA** unit is independent of the others and a form of conflict resolution based on priorities is applied to determine the final selected behavior. In the next stage, Signs act as IRM-like releasing enablers to control the expression of sub-sets of the animal's overall external behavior. Maes' (1991) architecture, with its explicit three-part (**SAS**) description of the structure of a static policy map, with explicit motivator links, but without explicit hierarchical control and without a proposed learning mechanism, acts as a bridge between the static and fully dynamic approaches to policy map construction.

In the context of this analysis, Catania's (1988) description of the operant three-part contingency (section 2.4), described in the light of this formulation, looks suspiciously like Tolman's (1932) *Sign-Gestalt Expectancy* (section 2.5), an explicitly anticipatory three-part Sign-Action-Sign (i.e. **SAS**) link. It seems unlikely that Skinner, as a staunch behaviorist, would have approved (Bjork, 1993). This level of analysis has identified and draws attention to the distinct possibility that these two largely antagonistic schools of thought are described by the same underlying phenomena; apparently only separated by a choice of different experimental conditions applied to the animal (i.e. the Skinner box vs. maze running) and, perhaps more significantly, with diametrically opposing interpretations on the results from those experiments. Of course, the results and outcomes of these experimental procedures from both traditions are not affected by this new treatment, but they are now unified into a single interpretation.

Much remains that can be done; the Action Selection Calculus lays a ground plan, but the devil remains in the detail. There exists a truly vast back catalogue of experimental data from the last 100 years of investigations that might be revisited in the light of this framework. Two substantive questions remain: (i) whether the link types, propagation and learning rules presented sufficiently describe the five factor theories and (ii) whether, even taken together as a whole, the five factor theories are sufficient to explain all animal behavior.

On the first, the theories are based on these experiments and much falls into place as a consequence. On the second, it seems unlikely – as evolutionary pressure has lead to incredibly diverse behavior patterns and mechanisms. The widespread nature of these experimental forms testifies to their recurring significance in the overall development of behavior. No single neural mechanism can or need be postulated, rather these strategies appear to have reoccurred frequently in different guises, in different species and different parts of the brain. Perhaps as a result of gradual evolutionary pressure, but equally because the strategy fits many different circumstances, much as mechanisms of *evolutionary convergence* are conjectured to recreate different physical forms with similar functions (e.g. Nishikawa, 2002).

11 Summary and Conclusions

This paper has presented a high-level view of the action selection properties of five central theories of natural action selection and learning and combined them into a single Action Selection Calculus. Each of these theories holds that actions are selected on the basis of prevailing sensory conditions. They do not agree on how this occurs, yet it is clear that each of the factor theories account for only a part of an individual animal's total behavioral repertoire and that what the experimenter observes is at least partly due to the design of their experiments. The paper has developed a set of five propagation rules and four learning strategies over three connection types to encapsulate and unify these otherwise apparently disparate approaches.

In conclusion, the Action Selection Calculus sets out to achieve several related goals. First, in providing a concise but comprehensive semi-formal definition of the various processes that go to make up the different behavioral and learning components that comprise the complete behavioral

repertoire of an animal. Second, the Calculus provides a viable structure with which to address one of the perennial problems in describing the different theories of animal learning – how to present the various theories as anything other than independent research topics. Third, it explicitly attempts to provide a framework with which to consider the functional relationships between the different factor parts, particularly in relation to the integration of innate (static) and learned (dynamic) policies. Fourth, the Calculus makes strong assertions about the sign anticipatory nature of the learning process. It is hoped that these will be considered by the neurological research community in parallel with more conventional reward based investigations (e.g. Schultz, 1998; Dayan and Balleine, 2002). Fifth, the methods of anticipatory learning encapsulated by the learning rules **L2**, **L3** and **L4** are clearly distinct from, but related to methods of reinforcement learning. While the discussion in this paper has been motivated from observations of animal behavior and tested empirically with a computer simulation the parallels with machine learning techniques are clear. It is therefore hoped that the same methods of analytic analysis will be applied to this class of algorithm as have been applied to reinforcement and neural network learning. Lastly, the Action Selection Calculus provides Agent and Animat researchers with a design specification for behavior selection methods, founded directly on a broad interpretation of natural learning and action selection theories.

APPENDIX ONE

Table of notational symbols used

*** TABLE 1 ABOUT HERE ***

APPENDIX TWO

The SRS/E Implementation

SRS/E acts as a benchmark implementation for the principles expressed in the Action Selection Calculus. The Action Selection Calculus is largely presented in implementation independent terms and might equally be coded directly in the style of a production rule system (e.g. Jones and Ritter, 2003; Klahr, Langley and Neches, 1987) or artificial neural network (e.g. Bishop, 1995). Other styles of implementation would highlight other aspects of the behavioral action selection and learning issues discussed and are to be welcomed.

Signs, Actions and the Link types are encoded directly as object data structures and stored in indexed and linked lists: the *sign_list* (**S**), *action_list* (**A**), *behavior_list* (**SA**) and *hypothesis_list* (for **SS** and **SAS**). Note that in SRS/E, **SS** and **SAS** links are considered as "micro-hypotheses" (*μ-hypotheses*), reflecting the notion that they each encapsulate a small, testable, predictive hypothesis about the structure of the environment. New links may be added to the lists at any time by learning. Each list object stores the parameters (activations {0, 1}, valences, weights, corroboration values, etc.) associated with its element and these are updated according to the principles of the propagation and learning rules presented here.

Execution cycle description

The execution of the system is based on a recurring cycle, each of which is considered to be one time step. The propagation and learning rules are applied to links in each list at each cycle. A detailed graphical representation of the SRS/E execution cycle is given in (Witkowski, 2003). The ordering of operations within the cycle is significant. At the start of the cycle, the activation values of the Signs are updated by interrogating the environment. Signs may be considered as *detectors*, in SRS/E Sign activation values are 0 or 1 (detected or not detected). Next, previously recorded predictions (in the

prediction_list) are compared to the current Sign activations and the anticipatory learning rules (**L2** and **L3**) applied and **SS** and **SAS** corroboration values *c* are adjusted. **SS** links in the *hypothesis_list* are sub_activated according to rule **P2**. **SA** links in the *behavior_list* are evaluated by rule **P1** (*direct_activation*) and candidate actions selected – this is the evaluation of the static policy map. Valence (goal) setting is considered a type of behavior in SRS/E (but may be changed by experimenter intervention) and Sign valences are updated if indicated by activations in the *behavior_list*.

If any Signs currently have valence, the top-goal *v'* is selected and a dynamic policy map constructed by repeated propagation of valence and sub_valence (by rule **P4**, *valence transfer*), until no more propagation is possible. Next, Rule **P5** (*valence_activation*) is applied to determine any candidate actions for activation from the DPM, comparing active Signs from the *sign_list* to sub-valenced Signs in the *hypothesis_list*. The Action associated with the **SAS** link in the DPM with the highest valence value and a currently active Sign is selected. The highest weight value *w* from the candidate actions from the SPM is compared to the valence value of the top goal (*v'*) at a subsumption point and the winning Action sent to the motor sub-system or actuators for final expression. When no action is available for expression, a default, low priority, exploratory action is selected (typically at random) from the *action_list*. Provision is made in the current SRS/E implementation to select default actions on the basis of guided exploration (e.g. *prioritized sweeping*, Moore and Atkeson, 1993) to improve search performance. Once the expressed Action has been chosen, rule **P3** (*Prediction*) may be applied to all **SS** and **SAS** rules and the *prediction_list* updated for evaluation in future cycles. Finally, once the Action has been expressed rules **L1** (*reinforcement rule*) updating the static policy map and **L4** (*anticipatory link formation*) creating new **SS** and **SAS** links to be added to the *hypothesis_list*, may be applied. The cycle then repeats.

End-notes:

[1] An action selection *policy map* may be viewed as *a set of situation – action rules for each of many states or sensed conditions*. Policy maps differ from planning methods (Russell and Norvig, 2003, sections II and IV), which typically deliver a prescribed sequence of actions leading to a desired situation (but see Schoppers, 1987, for a combined method).

[2] In this paper, a *static policy map* persists throughout the lifetime of the animal or animat, whereas *dynamic policy maps* are created in response to some immediate need and are transient.

[3] Note that the use of the term "S-R Behaviorism" here is distinct from "behaviorist school", which promoted the idea that all behavior of any significance could be observed, recorded and measured. All the approaches described here fall broadly into this category, which is not without its critics and in the case of human behavior seemingly with considerable justification (Velmans, 2000, for instance). The terms "associationist" and "cognitive" are likewise heavily over-loaded and are used with restricted historical meaning here.

[4] RL techniques don't, of course, actually stipulate that reward must be external, but external (food) reward is clearly indicated in this experimental procedure as the source of drive and motivation.

Acknowledgements

The author is grateful to Frank Ritter and Josh Gross, to Joanna Bryson, and to the anonymous reviewers of various drafts of this paper for their detailed and thoughtful comments. This work has been supported in part by EPSRC grant EP/C530683/1.

References

Agre, P.E. (1995). Computational Research on Interaction and Agency, *Artificial Intelligence*, **72**:1-52.

Baerends, G.P. (1976). The Functional Organization of Behaviour, *Animal Behaviour*, **24**:726-738.

Balkenius, C. and Morén, J. (1998) Computational Models of Classical Conditioning: A Comparative Study, *5th Int. Conf. on Simulation of Adaptive Behavior (SAB-5)*, pp. 348-353.

Barnett, S.A. (1975). *The Rat: A Study in Behavior*, Chicago: University of Chicago Press.

Barto, A.G. (1995). Adaptive Critics and the Basal Ganglia, in: Houk, J.C., Davis, J. and Beiser, D. (eds.) *Models of Information Processing in the Basal Ganglia*, pp. 215-232.

Barto, A.G. and Sutton, R.S. (1982). Simulation of Anticipatory Responses in Classical Conditioning by a Neuron-like Adaptive Element, *Behavioral Brain Research*, **4**:221-235.

Bellman, R. (1957). *Dynamic Programming*, Princeton, NJ: Princeton University Press.

Bi, G-Q. and Poo, M-M. (2001). Synaptic Modification by Correlated Activity: Hebb's Postulate Revisited, *Ann. Review of Neuroscience*, **24**:139-166.

Bindra, D. (1972). A Unified Account of Classical Conditioning and Operant Training, in Black, A.H. and Prokasy, W.F. (eds.) *Classical Conditioning II: Current Theory and Research*, New York: Appleton-Century-Crofts, pp. 453-481.

Bindra, D. and Stewart, J. (eds.) (1966). *Motivation*, Harmondsworth: Penguin Books Ltd.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press.

Bjork, D.W. (1993). *B.F. Skinner: A Life*, New York: Basic Books.

Blackman, D. (1974). *Operant Conditioning: An Experimental Analysis of Behaviour*, London: Methuen & Co.

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science*, Oxford University Press.

Bower, G.H. and Hilgard, E.R. (1981). *Theories of Learning*, Englewood Cliffs: Prentice Hall Inc., (citations used refer to the 1981 fifth edition).

Brooks, R.A. (1986). A Robust Layered Control System For a Mobile Robot, *IEEE Journal of*

Robotics and Automation, **RA2-**1:14-23.

Brooks, R.A. (1991a). Intelligence Without Reason, *MIT AI Laboratory, A.I. Memo No. 1293*. (Prepared for Computers and Thought, IJCAI-91, pre-print), April, 1991.

Brooks, R.A. (1991b). Intelligence Without Representation, *Artificial Intelligence*, **47**:139-159.

Bryson, J.J. (2000). Hierarchy and Sequence vs. Full Parallelism in Action Selection, *6th Int. Conf. on Simulation of Adaptive Behavior (SAB-6)*, pp. 147-156.

Bryson, J.J., Prescott, T.J. and Seth, A.K. (eds.) (2005). *Modelling of Natural Action Selection: Proceedings of an International Workshop*, AISB Press.

Butz, M.V., Sigaud, O. and Gérard, P. (2003). Internal Models and Anticipations in Adaptive Learning Systems, in: Butz, M.V., Sigaud, O. and Gérard, P. (eds), *Anticipatory Behavior in Adaptive Learning Systems*, LNAI-2684, pp. 86-109.

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

Courville, A.C., Daw, N.D. and Touretzky, D.S. (2006). Bayesian Theories of Conditioning in a Changing World, *Trends in Cognitive Sciences*, **10-**7:294-300.

Dayan, P. and Balleine, B.W. (2002). Reward, Motivation, and Reinforcement Learning, *Neuron*, **36**:285-298.

Hafting, T., Fyhn, M., Molden, S., Moser, M-B. and Moser, E.I. (2005). Microstructure of a Spatial Map in the Entorhinal Cortex, *Nature*, **436-**11:801-806.

Hallam, B.E., Hallam, J.C.T. and Halperin, J.R.P. (1994). An Ethological Model for Implementation in Mobile Robots, *Adaptive Behavior*, **3-**1:51-79.

Hart, P.E., Nilsson, N.J. and Raphael, B. (1968) A Formal Basis for the Heuristic Determination of Minimum Cost Paths, *IEEE Transactions on Systems Science and Cybernetics*, **SSC-4**(2):100-107.

Howard, R.A. (1960). *Dynamic Programming and Markov Processes*, Cambridge, MA: MIT Press.

Hebb, D.O. (1949). *The Organization of Behavior*, John Wiley & Sons.

Hergenhahn, B.R. and Olsen, M.H. (2005). *An Introduction to Theories of Learning*, Englewood Cliffs, N.J. Prentice-Hall (seventh edition).

Hubel, D.H. and Wiesel, T.N. (1962). Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex, *J. Physiology*, **160**:106-154.

Hull, C. (1943). *Principles of Behavior*, New York: Apple-Century-Crofts.

Jones, G. and Ritter, F.E. (2003). Production Systems and Rule-based Inference, in Nadel, L. (ed.) *Encyclopedia of Cognitive Science*, London: Nature Publishing Group, pp. 741-747.

Kaas, J.H. (1997). Topographic Maps are Fundamental to Sensory Processing, *Brain Research Bulletin*, **44**-2:107-112.

Kaelbling, L.P., Littman, M.L. and Moore, A.W. (1996). Reinforcement Learning: A Survey, *J. Artificial Intelligence Research*, **4**:237-285.

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A. and Guillot, A. (2005). Actor-Critic Models of Reinforcement Learning in the Basal Ganglia: From Natural to Artificial Rats, *Adaptive Behavior*, **13**-2:131-148.

Kirsh, D. (1991). Today the Earwig, Tomorrow Man? *Artificial Intelligence*, **47**:161-184.

Klahr, D., Langley, P. and Neches, R. (1987). *Production System Models of Learning and Development*, MIT Press.

Langley, P. (1996). *Elements of Machine Learning*, San Francisco, CA: Morgan Kaufmann

Leergaard, T.B., Alloway, K.D., Mutic, J.J. and Bjaalie, J.G. (2000). Three-Dimensional Topography of Corticopontine Projections from Rat Barrel Cortex: Correlations with Corticostriatal Organisation, *J. Neuroscience*, **20**-22:8474-8484.

Lorenz, K.Z. (1950). The Comparative Method in Studying Innate Behaviour Patterns, in: *Physiological Mechanisms in Animal Behaviour, Symposium of the Society of Experimental Biology*, Vol. 4, London: Academic Press, pp. 221-268.

MacCorquodale, K. and Meehl, P.E. (1953). Preliminary Suggestions as to a Formalization of Expectancy Theory, *Psychological Review*, **60**-1:55-63.

MacCorquodale, K. and Meehl, P.E. (1954). Edward C. Tolman, in: Estes, W.K. *et al*., (eds.) *Modern Learning Theory: A Critical Analysis of Five Examples*, New York: Appleton-Century-Crofts, pp. 177-266.

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

Maslow, A.H. (1987). *Motivation and Personality*, Third edition, Frager, R., *et al*. (eds.), New York: Harper and Row (first published 1954).

Minai, A.A. (1997). Covariance Learning of Correlated Patterns in Competitive Networks, *Neural Computation*, **9**:667–681.

Moore, A.W. and Atkeson, C.G. (1993). Prioritized Sweeping: Reinforcement Learning with Less Data and Less Time, *Machine Learning*, **13**:103-130.

Mowrer, O.H. (1956). Two-factor Learning Theory Reconsidered, with Special Reference to Secondary Reinforcement and the Concept of Habit, *Psychological Review*, **63**:114-128.

Nilsson, N.J. (1971). *Problem-solving Methods in Artificial Intelligence*, New York: McGraw-Hill.

Nishikawa, K.C. (2002). Evolutionary Convergence in Nervous Systems: Insights from Comparative Phylogenetic Studies, *Brain Behavior and Evolution*, **59**:240-249.

Oja, E. (1992). Principal Components, Minor Components, and Linear Neural Networks, *Neural Networks*, **5**:927-935.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*, Oxford: Clarendon Press.

Pavlov, I.P. (1927). *Conditioned Reflexes*, Oxford University Press (available as Dover Publications facsimile reprint, 1960).

Penfield, W. and Rasmussen, T.B. (1950). *The Cerebral Cortex of Man*, New York: Macmillan.

Razran, G. (1972). *Mind in Evolution: An East-West Synthesis of Learned Behavior and Cognition*, Boston: Houghton Mifflin Company.

Rescorla, R.A. and Wagner, A.R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement, in Black, A.H. and Prokasy, W.F. (eds.) *Classical Conditioning II: Current Theory and Research*, New York: Appleton-Century-Crofts, pp. 64-99.

Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*, Prentice Hall (citations used refer to the 2003 second edition).

Saksida, L.M., Raymond, S.M. and Touretzky, D.S. (1997). Shaping Robot Behavior Using Principles from Instrumental Conditioning, *Robotics and Autonomous Systems*, **22**-3/4:231-249.

Schmajuk, N.A. (1994). Behavioral Dynamics of Escape and Avoidance: A Neural Network Approach, *3rd Int. Conf. on Simulation of Adaptive Behavior (SAB-3)*, pp. 118-127.

Schoppers, M.J. (1987). Universal Plans for Reactive Robots in Unpredictable Environments, *Proc. 10th IJCAI*, pp. 1039-1046.

Shultz, W. (1998). Predictive Reward Signal of Dopamine Neurons, *J Neurophysiol*, **80**:1-27

Sejnowski, T.J. (1977). Storing Covariance With Nonlinearly Interacting Neurons, *J. Math. Biology* **4**:303-321.

Sejnowski, T.J. (1999). The Book of Hebb, *Neuron*, **24**-4:773-776.

Sejnowski, T.J., Dayan, P. and Montague, P.R. (1995). Predictive Hebbian Learning, *Proc 8th Computational Learning Theory (COLT-95)*, pp. 15-18.

Sewards, T.V. and Sewards, M.A. (2003). Representations of Motivational Drives in Mesial Cortex, Medial Thalamus, Hypothalamus and Midbrain, *Brain Res. Bulletin*, **61**:25-49.

Shettleworth, S.J. (1975). Reinforcement and the Organisation of Behavior in Golden Hamsters: Hunger, Environment, and Food Reinforcement, *J. Experimental Psychology: Animal Behavior Processes*, **104**-1:56-87.

Stolzmann, W., Butz, M.V., Hoffmann, J. and Goldberg, D.E. (2000). First Cognitive Capabilities in the Anticipatory Classifier System, *6th Int. Conf. on Simulation of Adaptive Behavior (SAB-6)*, pp. 287-296.

Sutton, R.S. (1988). Learning to Predict by the Methods of Temporal Differences, *Machine Learning*, **3**:9-44.

Sutton, R.S. (1991). Reinforcement Learning Architectures for Animats, *Int. Conf. on Simulation of Adaptive Behavior (SAB)*, pp. 288-296.

Sutton, R.S. and Barto, A.G. (1987). A Temporal-Difference Model of Classical Conditioning, *9th Ann. Conf. of the Cognitive Science Society*, pp. 355-378.

Sutton, R.S. and Barto, A.G., (1990). Time-Derivative Models of Pavlovian Reinforcement, in:

Gabriel, M. and Moore, J. (eds.) *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, MIT Press, pp. 497-537.

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press.

Thistlethwaite, D.A. (1951). Critical Review of Latent Learning and Related Experiments, *Psychological Bulletin*, **48**-2:97-129.

Tinbergen, N. (1951). *The Study of Instinct*, Oxford: Clarendon Press.

Thorndike, E.L. (1898). Animal Intelligence: An Experimental Study of the Associative Processes in Animals, *Psychol. Rev., Monogr. Suppl.*, **2**-8.

Tolman, E.C. (1932). *Purposive Behavior in Animals and Men*, New York: The Century Co.

Tolman, E.C. and Honzik, C.H. (1930). Introduction and Removal of Reward, and Maze Performance in Rats, *Univ. of California Publ. Psychol.* **4**:257-275, (cited Bower and Hilgard, 1981, pp. 335-338).

Travers, M. (1989). Animal Construction Kits, in Langton, C. (ed.) *Artificial Life; SFI Studies in Sciences of Complexity*, Reading, MA: Addison-Wesley, pp. 421-442.

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

Tyrrell, T. (1994). An Evaluation of Maes's Bottom-up Mechanism for Behavior Selection, *Adaptive Behavior*, **1**-4:387-420.

Velmans, M. (2000). *Understanding Consciousness*, London: Routledge.

Vogel, E.H., Castro, M.E. and Saavedra, M.A. (2004). Quantitative Models of Pavlovian Conditioning, *Brain Research Bulletin*, **63**:173-202.

Watkins, C.J.C.H. and Dayan, P. (1992). Technical Note: *Q*-learning, *Machine Learning*, **8**:279-292.

Welker, C. (1976). Receptive Fields of Barrels in the Somatosensory Neocortex of the Rat, *J. Comp. Neurology*, **166**-2:173-189.

Witkowski, M. (1997). Schemes for Learning and Behaviour: A New Expectancy Model, *Ph.D. Thesis*, University of London.

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

Witkowski, M. (2000). The Role of Behavioral Extinction in Animat Action Selection, *proc. 6th Int. Conf. on Simulation of Adaptive Behavior (SAB-6)*, pp. 177-186.

Witkowski, M. (2003). Towards a Four Factor Theory of Anticipatory Learning, in Butz, M.V., Sigaud, O. and Gérard, P. (Eds.) *Anticipatory Behavior in Adaptive Learning Systems*, Springer LNAI 2684, pp. 66-85.

List of Tables and Figures

Table 1

Symbol	Description
S	A *Sign*, detecting a specific condition
A	An *Action*, expressible by the agent
$\wedge$	Conjunction (AND), read as "coincident with"
$\vee$	Disjunction (OR)
${}^{w}\!\rightarrow_{t\pm\tau}$	Causal (S-R) link connection in an **SA** link
${}^{v,\,c}\!\leftrightarrows_{t\pm\tau}$	Anticipatory link connection in an **SS** or **SAS** link
ac(A)	The action cost of performing Action A (ac(A)≥1, by definition)
$t\pm\tau$	Cross-link time delay (t) and range (τ)
w	Weight, strength of connection in an **SS** link
c	Corroboration value associated with **SS** or **SAS** link
v(S)	Current *valence level* of Sign S
v'(S)	Valence level of the current *top-goal* (S)
α	The *corroboration rate*, update factor for c for a successful prediction
β	The *extinction rate*, update factor for c for a failed prediction
λ	*Learning rate*, probability that a learning rule **L4** activation will result in a new **SS** or **SAS** link

Figure 1

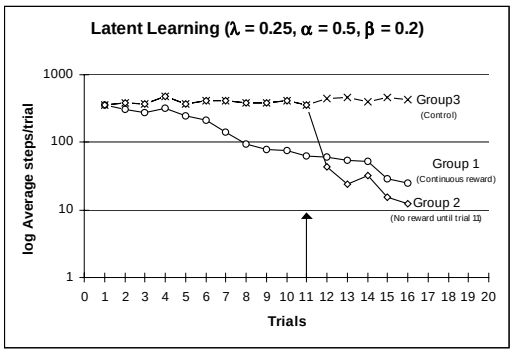

Latent Learning (λ = 0.25, α = 0.5, β = 0.2)
log Average steps/trial
1000
100
10
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Trials
Group3
(Control)
Group 1
(Continuous reward)
Group 2
(No reward until trial 11)

Figure 2

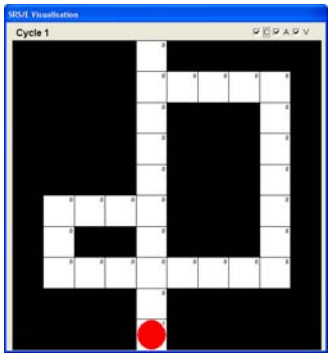


Figure 3

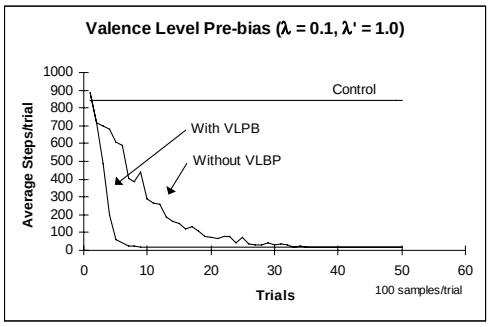
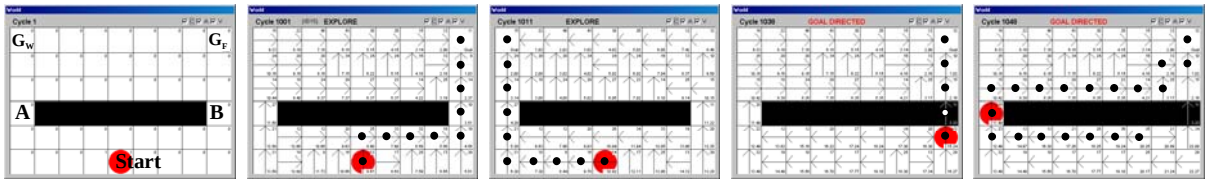

Valence Level Pre-bias (λ = 0.1, λ' = 1.0)
Average Steps/trial
1000
900
800
700
600
500
400
300
200
100
0
Control
With VLPB
Without VLBP
0
10
20
30
40
50
60
Trials
100 samples/trial

Figure 4



a b c d e

Mark Witkowski holds a B.Sc. in Biology with Computer Science with first class honours (1974) and a Ph.D. in Computer Science (1997), both from the University of London. He was a founding member of the Queen Mary College Artificial Intelligence and Robotics Laboratory and subsequently worked for Texas Instruments developing and managing networking products and manufacturing systems technologies. Witkowski joined Imperial College London in 1998 as a Research Fellow, researching in the areas of cognitive robotics, software agent technologies and modelling natural learning and behavior. Witkowski was programme chair for Towards Autonomous Robotic Systems (TAROS-06).
Address: Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ.
E-Mail: m.witkowski@imperial.ac.uk