Combining Action Selection Models with a Five Factor Theory

Mark Witkowski
Intelligent Systems and Networks Group
Department of Electrical and Electronic Engineering
Imperial College
Exhibition Road, London, SW7 2BT, United Kingdom
m.witkowski@imperial.ac.uk

Abstract

This paper describes a unifying framework for five highly influential but disparate theories (the five factors) of natural learning and behavioral action selection. These theories are normally considered independently, with their own experimental procedures and results. The framework 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. Exemplar experimental procedures will be used to discuss the areas of genuine difference, and to identify areas where there is overlap and where apparently disparate findings have a common source. 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.

1 Introduction

The overall aim of this paper is to provide a unifying description to encompass and combine five classical and highly influential "theories" of natural action selection and learning. These are the five factor theories. Each held a dominant place in theorizing during the 20th century and was supported by a wealth of meticulously gathered experimental data, but there has been little or no attempt to provide a single framework with which to rationally consider how they might interact.

The problem, in part, arises from the fact that these theories have been treated as largely competitive, at times with considerable animosity being generated between proponents of the differing approaches, or, more often, a tacit isolationism between the different schools of thought.

Such isolationism is surprising, as it clear that individual animals will demonstrate a whole range of behavioral phenomena, each of which might be most satisfactorily described by one or another of the approaches, largely depending on the circumstances the animal finds itself in. It is also very apparent that no single approach explains all animal action selection behavior.

Each factor theory is characterized by the underlying assumption that immediately 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 the principles by which that interaction was to be characterized. In itself, expressed behavior gives little 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, 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. This paper identifies where these mechanisms clearly differ, and where they are apparently different, but can be explained as manifestations of a single type of mechanism, and how these differences may be resolved into a single, structured framework. Given the range and diversity between individual animals and species, there is a fine balance to be struck between highly specific, quantitative, descriptions, trivially refuted due to this natural variation - and untestable generality. This paper attempts such a balance.

The five factor approach described here substantially extends, details and revises the approach to anticipatory learning and behavioral action selection introduced in [Witkowski, 2003]. The approach will be developed in the light of the *Dynamic Expectancy Model* (DEM) [Witkowski, 1998, 2000, 2003] and its actual (C++) computer implementation SRS/E. The analysis in this paper will be performed mainly at the level of the five factor theories, each of which is itself a digest of many exemplar experimental procedures. The paper will call on specific procedures where necessary, and illustrate issues with reference to the DEM and its implementation.

Section 2 provides a thumbnail sketch of each of the five factor theories. Comprehensive descriptions of the five theories can be found in any textbook of natural learning theory (e.g. [Bower and Hilgard, 1981]). Section 3 considers the 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 factor theories in the light of these component parts, and emphasizes the role of the action selection policy map, which may be either static or dynamic. Section 8 describes an arbitration mechanism between these policy maps, leading to final action expression.

2 The Five Factor Theories

The first of the five factor theories takes the form of *Stimulus-Response (S-R) Behaviorism*; which holds that action (the "response") selection is determined by the current sensory condition (the "stimulus"). Although first proposed in the final years of the 19th century [Thorndike, 1898], the approach continues to find contemporary support in the work of [Brooks, 1991; Bryson, 2000; and Maes, 1991]. This behavior is not defined by degree. The stimulus-response unit could be as apparently simple as a low-level reflex, such as the blink of an eye in response to a puff of air. Alternatively, behavioral repertories of considerable complexity can be postulated from essentially reactive models [Tyrrell, 1993; Tinbergen, 1951]. Such behaviors are generally considered to be innate (genetically determined) to the individual. Learning in the behaviorist regime is reward based, strengthening or weakening the connection between stimulus and response. It may be conjectured that not all such behaviors will be amenable to learning at the same rate, if at all.

The second factor theory, *classical conditioning*, was proposed by Ivan Pavlov (1849-1936) following observations that some innate reflexes can be associated with an otherwise neutral stimulus by repeated pairing, which will in turn elicit the reflex action. The procedure is highly repeatable and is easily demonstrated across a wide range of reflexes and species, and has been extensively modeled both mathematically and by implementation (e.g. [Vogel *et al*., 2004], for recent review).

The third theory, *operant* or *instrumental conditioning*, proposed by B.F. Skinner (1904-1990), who argued that actions were not "elicited" by impinging sensory conditions, but "emitted" by the animal in anticipation of a desired reward outcome. The effect is also highly repeatable under appropriate conditions, and it is clear that, given a suitable source of reward, an animal's (or indeed, a person's) behavior can be modified ("shaped") at will by judicious application of this principle. Whilst enormously influential in its time, only a relatively small number of computer models follow this approach (e.g. [Saksida *et al.*, 1997], or Schmajuk [1994] implementing Mowrer's [1956] "two-factor" theory, incorporating both classical and operant conditioning effects.)

The fourth theory, the "cognitive" model, proposed by E.C. Tolman [1932] describes a three-part basic cognitive unit, which establishes the expectation or anticipation of a specific stimulus following, and contingent on, an action taken in the immediate context of another stimulus. The context stimulus and action provide the means to achieve a desired and anticipated stimulus, the end. Tolman's *means-ends* approach both inspired and continues to be a fundamental technique of problem solving and planning for artificial intelligence ([Russell and Norvig, 1995], for instance). The Dynamic Expectancy Model (DEM) [Witkowski, 1998; 2000; 2003] and the Anticipatory Classifier System (ACS) model [Stoltzmann *et al*., 2000] represent recent three-part cognitive models.

A fifth theoretical position, broadly characterized by the term *associationism* (e.g. [Hebb, 1949]), concerns the direct associability and anticipation of stimuli following repeated pairing of activations. While of greater significance in other aspects of animal modeling, this approach does not directly incorporate an action component, and discussion of it will be restricted here to a minor supporting role in the action selection problem.

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 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 were necessary. The philosophically neutral term *sense data* might also be employed for this purpose (e.g. [Austin, 1962]). In the SRS/E model, S := {0,1}.

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*, (in SRS/E, by definition, $ac \geq 1$) 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 will be 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 highly differentiated or generalized sensing.

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 to other Signs holding valence.

As with activation and sub-activation, the valence and sub-valence properties may also be propagated between Signs under the conditions described in section 5. A Sign that is the direct source of valence is deemed *satisfied* once it has become active, and it and the propagated chain of sub-valenced Signs will revert to their normal, unvalenced, state (unless there are multiple sources of direct valence).

4 The Forms of Connection

The anticipatory stance proposes that the principal effects of the five target theories can be adequately explained 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. 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.

With the exception of a connection of type **C1**, 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 there are "bi-directional neurons", 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 theorems.

Connection type C1 (SA): $S_1 \overset{w}{\rightarrow}_{t\pm\tau} (A \wedge S_2)$

Connection type C2 (SS): $S_1 \overset{v,c}{\leftrightarrows}_{t\pm\tau} S_2$

Connection type C3 (SAS): $(S_1 \wedge A) \overset{v,c}{\leftrightarrows}_{t\pm\tau} S_2$

While connections of type **C1** have only an implicit anticipatory role, connection types **C2** and **C3** are both to be interpreted as making explicit anticipatory predictions.

The type **C1** connection ("SA") 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_2 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 **C2** 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 link has a *corroboration value*, *c*, associated with it, indicating the reliability of that prediction, based on continuing prior observation. A generic corroboration value update rule will be considered in section 6.1.

The *valence value*, *v*, of S_1 is a function of the current value of the valence value of S_2, and is hence 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. Where *t* is greater than zero (negative times have no interpretation in this context), the link is considered asymmetric. 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 **C2** connections.

The **C3** connection differs from **C2** 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. The transfer of valence *v* to S_1 needs to now be a function of both S_2 and the action cost 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.

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 "$\wedge$" should be interpreted as in section 4. On the right hand side of the rule, activate(), predict() and sub_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 any **C1** (SA) link,
if (active(S_1) $\vee$ sub_active(S_1))
then activate(A, *w*)

Rule P2 Sign Anticipation:
For any **C2** (SS) link,
if (active(S_1) $\vee$ sub_active(S_1))
then sub_active(S_2)

Rule P3 Prediction:
For any **C2** (SS) link,
if(active(S_1))
then predict(S_2, $t\pm\tau$)
For any **C3** (SAS) link,
if(active(S_1) $\wedge$ expressed(A))
then predict(S_2, $t\pm\tau$)

Rule P4 Valence transfer:
For any **C2** (SS) link,
if(valenced(S_2) $\vee$ sub_valenced(S_2))
then sub_valence(S_1, $f(v(S_2), d)$)
For any **C3** (SAS) link,
if(valenced(S_2) $\vee$ sub_valenced(S_2))
then sub_valence(S_1, $f(v(S_2), c, ac(A))$)

Rule P5 Valenced activation:
For any **C3** (SAS) link,
if(active(S_1) $\wedge$ sub_valenced(S_1))
then activate(A, *v'*)

Rule **P1** expresses the standard S-R behaviorist rule. Only in the simplest of animals would the activation of the action A lead to the direct overt expression of the action or activity. As there is no assumption that Signs are mutually exclusive, 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.

Rule **P2** allows for the propagation of 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. Evidence from second order classical conditioning studies would suggest that sub-activation propagates poorly (i.e. is heavily discounted).

Rule **P3** 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 **P4** indicates the spread of valence backwards along chains of anticipatory links. The sub_valence() process is shown in different forms for the **C2** (SS) and **C3** (SAS) links, reflecting the discounting (*d*) process mentioned earlier. As an exemplar, in the SRS/E model valence is transferred from S_2 to S_1 across the **C3** link according to the generic formulation: $v(S_1) := v(S_2) * (c / ac(A))$. By learning rule **L2** and **L3** (section 6.1) $0 < c < 1$, and as $ac(A) \geq 1.0$ (by definition), therefore $v(S_1) < v(S_2)$. Valence propagates preferentially across high confidence links with "easy" (i.e. lower cost value) Actions. Transfer is straightforward and has proved robust in operation in the DEM and SRS/E.

Rule **P5** indicates the activation of any Action A where the antecedent Sign S_1 is both active and valenced. As with rule **P1**, many Actions may be affected. The one associated with the highest overall S_1 valence value is selected.

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**. Note here that the valence value *v'* refers to the valence value of the Sign holding direct valence (the *top-goal*), whose value has been propagated to the SAS link, not that of either S_1 or S_2 of the **C3** (SAS) link in question.

6 The Forms of Learning

This section describes the conditions under which learning will take place. In the anticipatory 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 Sign. Each of the five factor theories takes a particular stance 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 characterised as advantageous/disadvantageous and thus interpreted as "rewarding" (or not) to the animal. This is frequently referred to as reinforcement learning. 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 **C3**, as described here, can be seen as subsuming link type **C1**, but the converse does not

hold. In the **C1** link, the role of anticipation in the learning process is implicit but is made explicit in the **C3** type link.

Learning rule L1 (the reinforcement rule):
For any **C1** (SA) link
if (active(A) $\wedge$ (valence(S_2) $\vee$ sub_valence(S_2)))
then update(*w*, α)

This is a generic form of the standard reinforcement rule. If the action is associated with any sensation that provides valence, then the connection weight *w* will be updated asymptotically by some factor α. Several well established weight update strategies are available, such as Watkins' *Q-learning* and Sutton's *temporal differences* (TD) method, see [Sutton and Barto, 1998] for review. 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 any (**C2** $\vee$ **C3**) link
if(predicted(S_2, $-t\pm\tau$) $\wedge$ active(S_2))
then update(*c*, α)

Learning rule L3 (anticipatory dis-corroboration):
For any (**C2** $\vee$ **C3**) link,
if(predicted(S_2, $-t\pm\tau$) $\wedge$ $\neg$active(S_2))
then update(*c*, β)

Learning rule L4 (anticipatory link formation):
if($\neg$predicted(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 **C2** and **C3** link types. Three conditions are significant, where a prediction has been made, and the predicted event did occur at the expected time (learning rule **L2**). The link is considered corroborated and is strengthened. Where a prediction is made, but the event does not occur (learning rule **L3**), the link is considered dis-corroborated and weakened. Lastly, where an event occurs, but it was not predicted at all (learning rule **L4**).

The SRS/E computer implementation employs the simple but robust, effective and ubiquitous update rule $c := c+\alpha\,(1-c)$, where $(0 \leq \alpha \leq 1)$ for **L2**, and the generic update rule $c := c-\beta(c)$, where $(0 \leq \beta \leq 1)$, is again simple, effective and robust for **L3**. Both update functions are asymptotic towards 1.0 and zero respectively. The net effect of these update rules is to maintain a form of "running average" more strongly reflecting recent outcomes, with older outcomes becoming successively discounted (tending to zero contribution). The greater the values of α and β, the more aggressively recent events are tracked. The particular settings of these values are specific to the individual animal. Where no prediction was made by a rule, *c* remains unchanged regardless of the occurrence of S_2. This is consistent with the notion that a rule is only responsible for predicting an event under the exact conditions it defines.

The key issue here is that anticipatory learning is *everytime*. Every prediction made, regardless of its cause, initiates learning. Learning is independent of valenced reward (this is the phenomenon of *latent learning* [Witkowski, 1998], [Thistlethwaite, 1951]). Anticipatory links are measured relative to their predictive ability, not their usefulness. Correct anticipation is its own reward. Such anticipatory reward is generated locally to the **C2** or **C3** link, and is independent of all others. Further, if circumstances change, each link adjusts automatically to the prevailing circumstances based on recent predictive experience. Anticipation may also be combined with valence, to preferentially focus the learning process on Signs that have, or have had, valence (e.g. the *Valence Level Pre-Bias* technique [Witkowski, 1998]).

Where an event is unpredicted by any link, this is taken as a cue to establish a new link between the unpredicted event (as S_2) and some recent recently active event (as S_1) at time *t*, rule **L4**. Where a **C3** link is created some expressed Action A contemporary with the new S_1 is also implicated. Again the choice of how many new links are formed, and the range of values for *t* and τ are specific to the individual animal. 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 learning rules **L2** and **L3** separate the effective from the ineffective.

The key issue here is that link learning may be invoked everytime a novel or unpredicted Sign is detected. Learning may proceed from *tabula rasa*, and is rapid while much is novel. In a restricted environment, link learning will slow as more is predicted, but resume if circumstances change.

No rule for link removal is considered here, but has been discussed elsewhere in the context of the 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].

7 Explaining the Five Factors

This section returns to the action selection factor theories outlined in section 2, and will discuss them 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, others require a combination of forms, and yet others perhaps require re-interpretation in the light of this formulation.

7.1 Stimulus-Response Behaviorism

S-R Behaviorism holds that all, or the majority of, observed and intelligent behavior can be ascribed to an innate, pre-programmed, pairing of sense data driven stimuli and pre-defined actions.

7.1.1 Static Policy Maps

With no embellishments, S-R behaviorism is reduced to connection type **C1** and propagation type **P1**. The underlying assumption that all these strategies adopt 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 indicated. We may refer to this as a *static policy map*. The DEM records these connections in a list, effectively ordered by the weight parameter, *w*. Recall that the weighting value *w* may be modified by reinforcement learning [Sutton and Barto, 1998].

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; Maes, 1991; Tinbergen, 1951; Tyrrell, 1993].

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, it can be 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 response* (UR) to the *unconditioned stimulus* (US): the innate reflex. As reflexes are typically unconditionally expressed (i.e. 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 **C2** 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, and is dynamic according to that history, by learning rules **L2** and **L3**, the rates depending on the function 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. We interpret this link 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 **C3** 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 **C3** (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**), and in the absence of any alternative, continue to do so. Action selection is now firmly contingent on both encountered Sign and prevailing valence.

Due to valence transfer (**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 **C3** contingency $(S_C \wedge A_M) \leftrightarrows S_L$ would provide propagated valence to S_C 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. This transfer of valence may be used to create long chains of behaviors (such as in preparing animals for film performances) by building the sequence back one step at a time.

Propagation rule **P4** also allows for *secondary* or *derived reinforcement* effects ([Bower and Hilgard, 1981], p.184), in which normally non-reinforcing **C2** links may be paired with (or even chained from) an innately valenced one.

7.4 Tolman's Expectancy Model

Catania's [1988] description of the operant three-part contingency, described in the light of this formulation looks suspiciously like Tolman's [1932] *Sign-Gestalt Expectancy*, an explicitly anticipatory three-part Sign-Action-Sign (i.e. **C3**) link. Skinner, as a staunch "old-school" behaviorist, would definitely not have approved! Where the Skinner box

investigates the properties of the individual **C3** 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 locations in the maze may be represented as Signs (a rat may only be in one location at once, though it may be mistaken as to which one), and traversal between them as identifiable Actions. Every location-move-location transition may be represented as an anticipatory **C3** 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, placed in a maze, the animal will learn the structure as a number of **C3** connections with or without (i.e. latently) the presence of valence or reward. Novel locations encountered by random (or guided) exploration invoke **L4**, and the confidence value *c* is updated each time a location is revisited, by **L2** or **L3**. Once encountered, food may impart valence to a location (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 **C3** (and indeed **C2**) links to establish a *dynamic policy map* (DPM). This takes the form of a graph of all reachable Signs. In SRS/E this is considered as form of a modified breadth first search, in which each Sign node is assigned the highest propagated valence value. Again this generic process, as implemented in SRS/E, is computationally fast and robust in operation.

Once created, each Sign implicated in the DPM is associated with a single Action from the appropriate **C3** link, the one on the highest value valence path, and a single valence value *v*, indicating its rank in the 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 7.3 looks to be no more than an attempt to manipulate the naturally constructed dynamic policy to prefer one chain of actions to all the others.

The dynamic policy map must be recomputed 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 1 illustrates this (from [Witkowski, 2000]). The animat (circle) is in a grid maze, and each square represents a location Sign, and the arrows indicate the current policy action in that square. The animat was allowed to explore the maze shown on the left completely by selecting random actions, but without any source of valence (i.e. latently). When G is given valence, the animat builds a DPM and takes the shortest path via B. With the animat returned to S, and G still valenced, but B now blocked, the DPM will still indicate a path via B (the blockage is undiscovered), center. As the intended (up) action to B now fails, the DPM alters to prefer the apparently longer path via A, and the observable behavior of the animat will abruptly change as a new DPM is constructed and a new path is preferred, right.

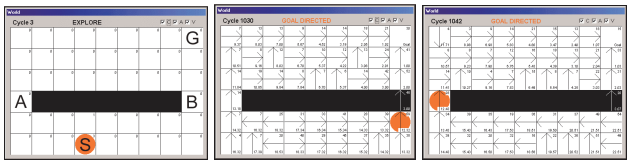

Figure 1: Rapid changes in the Dynamic Policy Map

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, and the valence driven dynamic policy map. The dynamic policy map is transient, and must interleave with the largely permanent static policy map. The valence value of the original source (*v'* from section 5, the top-goal) is (numerically) equated to the **C1** connection weight values, *w*. While $v' \geq \text{active}(w)$, actions are selected only from the DPM. If at any point $\text{active}(w) > v'$, DPM selection is suspended, and actions are taken from the static policy. Once completed or abandoned, control reverts to the DPM.

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

Static policy maps may also be partitioned. Tinburgen [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, with all its subsidiary SR connections simultaneously enabled, but then individually ranked within that grouping. Selection may then proceed as for the Dynamic Policy Map example. Note that in the DEM, valence setting is reserved as a static policy map activity, a type of Action. In this context the IRM releasing enablers start to look, in evolutionary terms, like the beginnings of valenced items.

9 Summary and Conclusions

This paper has presented a high-level view of the action selection properties of five central theories of behavior and learning. 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 it may be demonstrated experimentally that each theory accounts for only a part of an individual animal's behavioral repertoire, and that what the experimenter sees 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.

This has lead to the notion of different types of policy map operating within the animal, from static to dynamic, and 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.

Much remains to be done. This overview paper has laid 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 links, 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. Identifying these experimentally observed exceptions will serve to refine the multi-factor approach presented, leading in time to a better, more encompassing, solution.

Even though one can observe classical and operant conditioning, and means-ends behavior in humans, it is abundantly clear than even taken together the five factors fail to explain human behavior to a very considerable extent. It is vastly apparent that human (and possibly other primate) activities are not solely, or even predominantly, driven directly by immediately prevailing and observable circumstances. However, one might see these five mechanisms as both a foundation for, and a bridge to, the evolutionary development of higher-level cognitive functions.

References

[Austin, 1962] Austin, J.L. *Sense and Sensibilia*, Oxford University Press, 1962

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

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

[Bower and Hilgard, 1981] Bower, G.H. and Hilgard, E.R. *Theories of Learning*, Englewood Cliffs: Prentice Hall Inc., fifth edition, 1981

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

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

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

[Hebb, 1949] Hebb, D.O. *The Organization of Behavior*, John Wiley & Sons, 1949

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

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

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

[Russell and Norvig, 1995] Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

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

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

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

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

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

[Tinbergen, 1951] Tinbergen, N. *The Study of Instinct*, Oxford: Clarendon Press, 1951

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

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

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

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

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

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

[Witkowski, 2003] Witkowski, M. Towards a Four Factor Theory of Anticipatory Learning, in Butz, M.V. *et al.* (Eds.) Anticipatory Behavior in Adaptive Learning Systems, Springer LNAI 2684, pages 66-85, 2003