No Bad Dogs: Ethological Lessons for Learning in Hamsterdam*

Bruce M. Blumberg
MIT Media Lab
E15-305, 20 Ames St.
Cambridge Ma. 01463
bruce@media.mit.edu

Peter M. Todd
Max Planck Inst. for Psychological Research,
Center for Adaptive Behavior and Cognition
Leopoldstrasse 24, 80802 Munich Germany
ptodd@mpipf-muenchen.mpg.de

Pattie Maes
MIT Media Lab
E15-305, 20 Ames St.
Cambridge Ma. 01463
pattie@media.mit.edu

Abstract

We present an architecture for autonomous creatures that allows learning to be combined with action selection, based on ideas from ethology. We show how temporal-difference learning may be used within the context of an ethologically inspired animat architecture to build and modify portions of the behavior network, and to set fundamental parameters including the strength associated with individual Releasing Mechanisms, the time course associated with appetitive behaviors, and the learning rates to be used based on the observed reliability of specific contingencies. The learning algorithm has been implemented as part of the Hamsterdam toolkit for building autonomous animated creatures. When implemented in Silas, a virtual dog, the algorithm enables Silas to be trained using classical and instrumental conditioning.

1 Introduction

Action selection and learning represent two significant areas of research in behavior-based AI [Maes94], and advances in both areas are essential if we are to build animats that demonstrate robust adaptive behavior in dynamic and unpredictable environments. However, most work in this area to date has focused on one problem or the other. Several researchers [Maes90a, Blumberg94, Blumberg95, Tyrrell94, Tu94] have proposed ethologically inspired models of action selection that purport to handle many of the challenges associated with building animats which must juggle multiple goals at one time. However, this work did not incorporate learning, and the proposed architectures generally involved hand-built behavior networks and "not-a-few" parameters that had to be tweaked. In contrast, there has been a great deal of impressive work in learning [Watkins89, Whitehead92, Kaelbling92, Lin92, Mahadevan91, Sutton91], but the focus of this work has been on single-goal systems and on issues of optimality. Moreover, while it is now understood how to learn an optimal policy in certain kinds of worlds, it is also recognized that this is just the tip of the iceberg and that addressing the issues of "perceptual aliasing" and the "curse of dimensionality" are perhaps more important than the details of the underlying learning algorithm itself. In other words, the structure in which learning takes place is as critical as the way the learning actually occurs. Among the few projects that pull these two research strands together, Klopf [Klopf93] developed a model that accounts for a wide range of results from classical and instrumental conditioning, and Booker [Booker88] and Maes [Maes90b] have integrated learning into more complete action selection algorithms. But much remains to be done.

In this paper, our contribution is to show how certain types of associative learning may be integrated into the action selection architecture previously proposed by Blumberg [Blumberg94, Blumberg95]. The underlying learning algorithm we use is Sutton and Barto's Temporal Difference model [Sutton90], which builds and modifies portions of the behavior network in our ethologically inspired system. Furthermore, this algorithm sets several of the fundamental variables associated with our system, thus addressing some of the previous criticisms about proliferating free parameters. In particular, our system will:

- learn that an existing behavior can lead to the fulfillment of some previously-unassociated motivational goal when performed in a novel context (i.e., in more ethological language, the system will learn new appetitive behaviors).
- learn the context in which an existing behavior will directly fulfill a different motivational goal (i.e. learn new releasing mechanisms to be associated with a consummatory behavior).

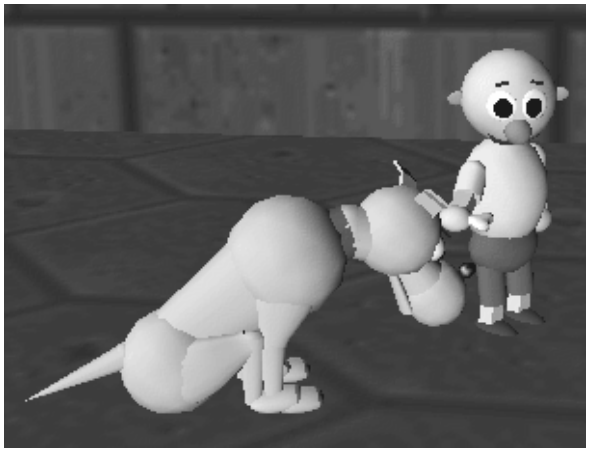

Figure 1: Silas T. Dog and his faithful teacher Dr. J.T. Puppet

* Apologies to Barbara Woodhouse [Woodhouse82].

- learn new releasing mechanisms that are found to be useful predictors of the future state of the world (i.e. show classical conditioning).

This approach to action selection combined with learning has been implemented and incorporated into the architecture for "Silas T. Dog," an autonomous, animated virtual dog featured in the ALIVE interactive virtual reality project [Maes95]. It has been successfully used to teach this old dog new tricks. By giving Silas the ability to learn, we hope to demonstrate the scientific feasibility of the coherent ethology-based model of animal behavior described in this paper. But this work also shows the usefulness of adding adaptability and user-trainability to autonomous VR agents intended to interact with people in an engaging and convincing manner. In this paper, we first discuss the important lessons from ethological research that have inspired the design of this learning system (Section 2), and then describe the basic action selection system underlying Silas's behavior (in Section 3). We proceed to the new learning mechanisms in Section 4, and indicate their impact on behavior in Section 5. Finally, we discuss the implications and uses of this approach in Section 6.

2 Lessons from Ethology

Like the previous work on action selection in Hamsterdam [Blumberg94], our present efforts to incorporate learning are inspired by decades of ethological research [Dickinson94, Gallistel90, Gallistel94, Gould94, Lorenz73, Sutton90, McFarland93, Plotkin93, Shettleworth94, Davey89]. Here we briefly list the main take-home messages we have received from this literature and used with Silas to date.

Learning complements evolution: Learning is a mechanism for adapting to significant spatial and temporal aspects of an animal's environment that vary predictably, but at a rate faster than that to which evolution can adjust (e.g. predictable changes occurring within a generation). The most adaptive course in this case is to evolve mechanisms that facilitate learning of these rapidly-varying features. Thus, evolution determines much of what can be learned and the manner in which it is learned. While this is sometimes viewed as imposing negative "constraints on learning," the innate structure in which learning occurs most often has the effect of dramatically simplifying the learning process [Gallistel90, Gallistel94, Gould94, Lorenz73, Miller90, McFarland93, Plotkin93, Shettleworth94, Todd91].

Motivations and goals drive learning: Animals learn things that facilitate the achievement of biologically significant goals. Thus, motivations for those goals drive learning. It is far easier to train a hungry animal using food as a reward than a satiated one [Gould94, Lorenz73, McFarland93]. The components of the lesson to be learned--that is, what things or actions facilitate a particular goal--are usually, but not always, contiguous in some sense with the satisfaction of the goal. For instance when a dog gets a cookie, the thing to learn is what behavior it performed just before getting that cookie.[1] (Taste-aversion learning occurring over a much longer time-period is an important example of non-temporally contiguous learning that we also wish to model.)

Animals learn new appetitive behaviors: Operant and instrumental learning can be viewed as an animal's discovery of new appetitive behaviors--that is, new behaviors that bring them closer to attaining some goal [Lorenz73]. Specifically, operant conditioning occurs when an animal finds that an existing behavior (one that was typically performed in another context for another purpose) performed in a new stimulus situation will reliably lead to a state of the world in which performance of a consummatory behavior will successfully satisfy one or more motivational variables. For example, a dog that learns to roll over on command in order to get a cookie has added the "roll over in the context of the spoken word ROLL" behavior to its appetitive behaviors for feeding.

Animals learn the value of stimuli and behaviors: Stimuli that identify motivationally significant contexts, and appetitive behaviors that lead to goal-satisfying contexts, both have learnable values associated with them. This is an underlying assumption of many associative models of classical or operant conditioning [Sutton90, Klopf93, Montague94]. It is also a necessary assumption for integrating learning into an ethologically inspired model, such as that used in Hamsterdam, in which external and internal factors are expressed as real-valued quantities.

Animals learn more than associations: It is not enough simply to learn that some action X is a useful appetitive behavior. It is also useful to learn an expected time course for X, that is, how long to engage in action X before concluding that reinforcement is not forthcoming and that another behavior may be more successful. Gallistel [Gallistel90,94] presents evidence that in certain cases animals can learn the information relevant to this kind of time course knowledge, including the number of events, the time of their occurrence and the interval between them. Within an associative model, such as the one we develop here, it is also necessary to augment the system's knowledge with such additional statistics in order to allow learning of a behavior's expected time course.

---

1 Credit assignment is a hard problem for both animals and machines. Many of the innate mechanisms which facilitate learning in animals do so by solving, in effect, the credit assignment problem. This is done by either making the animal particularly sensitive to the relevant stimuli, or predisposed to learn the lesson when it is easiest to do so (e.g. song learning) [Gould94, McFarland93, Plotkin93]. With respect to animal training, trainers stress the importance of giving reinforcement within 2-4 seconds of the performance of a desired behavior and it is very difficult to train animals to perform sequences of actions, except by chaining together individually trained actions [Rogerson92, Lorenz94]. However, by breaking a sequence of desired actions into individually trained steps, the trainers are solving the credit assignment problem for the animal. Learning and training are flipsides of the same coin: learning tries to make sense of the world and put it into as simple a framework as possible, and training tries to make the world as simple as possible for the learning mechanisms to grasp.

3 Computational Model

In this section, we summarize our ethologically inspired computational model for action selection and learning in autonomous animated creatures. The underlying architecture for motor control and action selection is described in more detail in [Blumberg94, Blumberg95]. Here we will only describe enough of the underlying architecture to make it clear how learning is integrated into this system.

3.1 Overview

The basic structure of a creature in our system consists of the three basic parts (Geometry, Motor Skills and Behavior System) with two layers of abstraction between these parts (Controller, and Degrees of Freedom).The Geometry provides the shapes and transforms that are manipulated over time to produce the animated motion. The Motor Skills (e.g. "walking," "wagging tail") provide atomic motion elements that manipulate the geometry in order to produce coordinated motion. Motor Skills have no knowledge of the environment or state of the creature, other than that needed to execute their skill. Above these Motor Skills sits the Behavior System, which is responsible for deciding what to do, given a creature's goals and sensory input. The Behavior System triggers the correct Motor Skills to achieve the current task or goal. Between these three parts are two layers of insulation, the Controller and the Degrees of Freedom (DOFs), which make this architecture generalizable and extensible. The DOFs, Controller, and Motor Skills together constitute what we call the Motor System.

A key feature of the Motor System is that it allows multiple behaviors to express their preferences for motor actions simultaneously. It is structured in such a way that Motor Skills that are complementary may run concurrently (e.g. walking and wagging the tail), but actions that are incompatible are prevented from doing so. Furthermore, the motor system recognizes 3 different imperative forms for commands: primary commands (i.e. do the action if the system is able to), secondary commands (i.e. do it if no other action objects) and meta-commands (i.e. do it this way). This allows multiple behaviors to express their preferences for motor actions.

There are at least three types of sensing available to autonomous creatures in our system:

- Real-world sensing using input from physical sensors.
- "Direct" sensing via direct interrogation of other virtual creatures and objects.
- "Synthetic vision" using computer vision techniques to extract useful information from an image rendered from the creature's viewpoint. This is used for low-level navigation and obstacle avoidance, based on ideas from Horswill [Horswill93], and [Reynolds87].

Most of the learning in our system relies on "direct" sensing, in which objects in Silas's world make information about their state available to Silas and other objects. For example, in the ALIVE interactive virtual reality system [Maes96], a user's gestures in the "real world" are converted into state variables associated with the user's proxy "creature" in Silas's world. Via direct sensing Silas can "sense" the state of this proxy creature, and thus react to the gestures of the user in the real world.

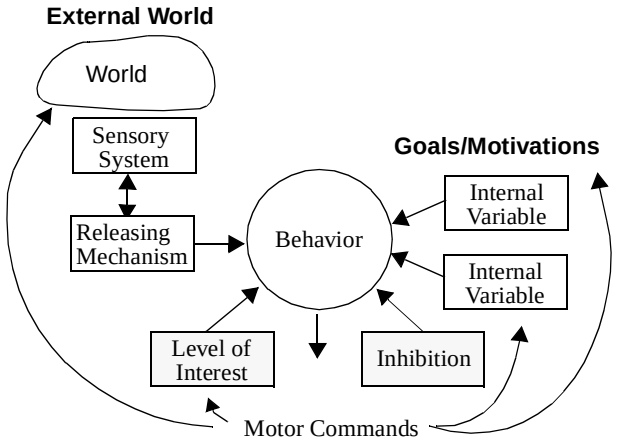

Figure 2: The purpose of a Behavior is to evaluate its relevance given external stimulus and internal motivations, and if appropriate issue motor commands. Releasing Mechanisms act as filters which identify significant objects or events from sensory input, and output a value which represents the strength of the sensory input. Motivations or goals are represented via Internal Variables which output values which represent their strength. A Behavior combines the values of the Releasing Mechanisms and Internal Variables on which it depends to yield its value before Level of Interest and Inhibition from other Behaviors. Level of Interest is used to model boredom or behavior-specific fatigue. Behaviors compete, via mutual inhibition, with other Behaviors for control of the creature.

The purpose of the Behavior System is to send the "right" set of control signals to the motor system at every time-step. The "right set" is determined by weighing the current competing goals of the creature, assessing the state of the environment, and choosing the behavior (and its associated set of actions) that best satisfies some of the creature's goals at this instant in time. More generally, the Behavior System coordinates the creature's use of its available high-level behaviors in a potentially unpredictable environment.

We will now turn to a discussion of the major components of the Behavior System and briefly describe the role of each component.

3.2 Behaviors

The Behavior System is a distributed network of self-interested, goal-directed entities called Behaviors. The granularity of a Behavior's goal can range from very general (e.g. "reduce hunger") to very specific (e.g. "chew food"). The major components of an individual Behavior are shown in Figure 2.

Each Behavior is responsible for assessing its own current relevance or value, given the present state of internal and external factors. On the basis of this assessed value, each Behavior competes for control of the creature. The current value of a Behavior may be high because it satisfies an important need of the creature (as indicated by high values of the associated Internal Variables--see section 3.4), or because its goal is easily achievable given the current state of the environment (as indicated by high values of the associated Releasing Mechanisms--see Section 3.3).

Behaviors influence the system in several ways: by issuing motor commands which change the creature's relation-

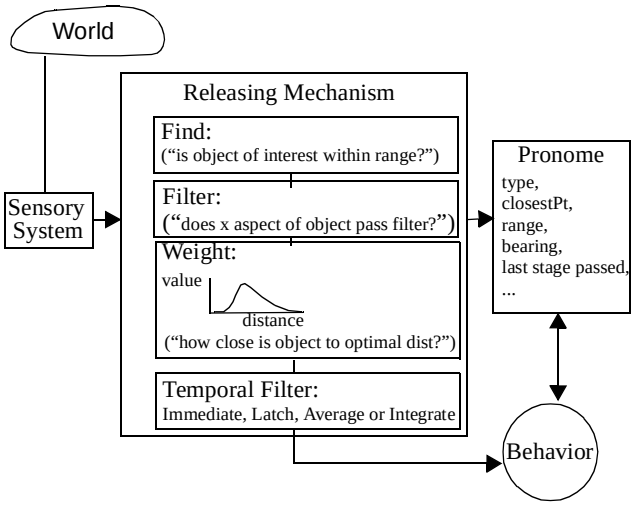

Figure 3: Releasing Mechanisms identify significant objects or events from sensory input and output a value which represents its strength. By varying the allowed maximum for a given Releasing Mechanism, a Behavior can be made more or less sensitive to the presence of the relevant stimuli. A Releasing Mechanism also fills in a data structure called a Pronome [Minsky86]. The Pronome can be thought of as representing the current focus of attention, which can be shared across behaviors.

ship to its environment, by modifying the value of Internal Variables, by inhibiting other Behaviors, or by issuing suggestions which influence the motor commands issued by other Behaviors.

Behaviors are distinguished from Motor Skills in two ways. For example, Behaviors are goal-directed whereas Motor Skills are not. For example, "Walking" is a Motor Skill whereas "Moving toward object of interest" is a Behavior which invokes the "Walking" Motor Skill to accomplish its goal. Second, Motor Skills do not decide when to become active, but rather rely on another entity to invoke them when appropriate. Motor Skills are similar to the Ethological idea of "Fixed Action Patterns" in that once initiated they do not require sensory input (other than proprioceptive input) and they have a specific time course [Lorenz76]. For the purposes of learning, the set of Motor Skills is assumed to be fixed. Thus, Motor Skill learning is an important type of learning which we do not address in this system at present. However, see [Giszter94] for an interesting example of motor learning which uses Maes' algorithm.

3.3 Releasing Mechanisms

Objects called Releasing Mechanisms (see Figure 3) filter the creature's sensory input to identify objects and events whose presence helps determine the current appropriateness of each possible Behavior. Releasing Mechanisms output a continuous value that typically depends on:

- the presence of the stimulus object or event (e.g. a person is nearby).
- more specific features associated with the stimulus (e.g. the person's hand is down and extended).
- the distance to the stimulus relative to some ideal distance (e.g. the hand is one foot away, and I like hands that are two feet away).

By representing the output of a Releasing Mechanism as a continuous quantity, this value may be easily combined with the strength of motivations (from Internal Variables) that are also represented as continuous values. This combination in turn allows the creature to display the kind of trade-off behavior one finds in nature where a weak stimulus (e.g. week-old pizza) but a strong motivation (e.g. very hungry) may result in the same behavior as a strong stimulus (e.g. chocolate cake) but weak motivation (e.g. full stomach).

Releasing Mechanisms return a value between some minimum and maximum, with the maximum value occurring when the Releasing Mechanism's object of interest appears at some optimal distance from the creature. But how can we set this maximum value for a given Releasing Mechanism? The value of the Releasing Mechanism should reflect the usefulness of its associated Behavior to the creature, and hence its maximum should be equal to the time-discounted usefulness of its object of interest (which the Behavior will exploit). For example, for a food-sensitive Releasing Mechanism, its maximum value should be equivalent to the time-discounted value of the food source to the creature. We use the time-discounted value to reflect the fact that it will take a finite period of time to consume the food. Thus, with everything else being equal, a food source that provides more energy per time-tick will be favored over one which provides less energy per tick, even if the total amount of energy from the two food sources is the same. This perspective is central to our approach to learning.

Ethologists generally view Releasing Mechanisms as filtering external stimuli. We generalize this and allow Releasing Mechanisms to be inward looking as well, so for example, a Releasing Mechanism might be sensitive to the performance of a given behavior. This latter type of Releasing Mechanism is called a Proprioceptive Releasing Mechanism (PRM) to distinguish it from External Releasing Mechanisms (ERM) which are outward looking.

3.4 Internal Variables

Internal Variables are used to model internal state such as level of hunger or thirst. Like Releasing Mechanisms, Internal Variables are each expressed as a continuous value. This value can change over time based on autonomous increase and damping rates. In addition, certain Behaviors, such as "eating," modify the value of an Internal Variable as a result of their activity. Ethologically speaking, these are "consummatory" behaviors, which when active have the effect of reducing the further motivation to perform that behavior. When a consummatory Behavior is active, it reduces the value of its associated motivational variable by an amount equal to some gain multiplied by the Behavior's value. Thus, the higher the Behavior's value, the greater the effect on the motivational Internal Variable.

Much of the learning that occurs in nature centers around the discovery either of novel appetitive behaviors (i.e. behaviors that will lead to a stimulus situation in which a consummatory behavior is appropriate), or of stimulus situations themselves in which a consummatory behavior should become active. More generally, one could say that the motivations that these consummatory behaviors satisfy drive much of the learning in animals. We adopt just such a perspective in our model: motivational Internal Variables are viewed as entities that actively drive the attempt to discover

new strategies (i.e. appetitive behaviors) that insure their satisfaction.

Based on this perspective, we can use the change in the value of a motivational variable due to the activity of a consummatory behavior as a feedback (or reinforcement) signal for the learning process. As we stated earlier in this section, changes in motivational Internal Variables are proportional to the value of the Behavior that changes them, which is in turn proportional to the value of the Releasing Mechanism that triggers it and the current level of motivation reflected in the Internal Variable value. Thus, more succinctly, motivational change-based reinforcement is proportional to both the quality of the stimulus (i.e., the value of the underlying Releasing Mechanism) and the level of motivation (i.e., the value of the changing Internal Variable).

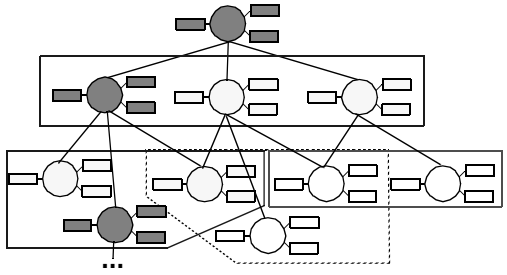

Figure 4:Behaviors are organized into groups of mutually inhibiting behaviors called Behavior Groups. Behavior Groups in turn are organized in a loose hierarchical fashion. Behavior Groups at the upper levels of the hierarchy contain general behaviors (e.g. "engage-in-feeding") which are largely driven by motivational considerations, whereas lower levels contain more specific behaviors (e.g. "pounce" or "chew") which are driven largely by immediate sensory input. The algorithm's arbitration mechanism insures that only one Behavior in a given Behavior Group will have a non-zero value after inhibition. This Behavior is then active, and may either issue primary motor commands, or activate the Behavior Group which contains its children behaviors (e.g. "search-for-food", "sniff", "chew" might be the children of "engage-in-feeding"). The dark gray nodes represent the path of active Behaviors on a given tick. Losing Behaviors in a given Behavior Group may nonetheless influence the resulting actions of the creature by issuing either secondary or meta-commands.

3.5 Behavior Groups

Behaviors are organized into mutually inhibiting groups called, simply, Behavior Groups (similar to Minsky's cross-exclusion groups [Minsky86]). These are illustrated in Figure 4. While we find a loose hierarchical structure useful this is not a requirement (i.e. all the Behaviors can be in a single Behavior Group). Behavior Groups are important because they localize the interaction among Behaviors which in turn facilitates adding new Behaviors.

Typically, the Behavior Group at the top of the system is composed of Behaviors that, in turn, have their own subsumed Behavior Groups specialized for addressing a particular need of the creature (e.g. feeding or mating). These subgroups very often have the same structure, consisting of a consummatory behavior and one or more appetitive behaviors. The different appetitive behaviors represent alternative strategies for bringing the creature into a situation in which the consummatory behavior may be used. Associated with each of the appetitive behaviors are one or more Releasing Mechanisms that signal the appropriateness of this behavior given the current environment (i.e., the likelihood that this appetitive behavior will lead to a situation in which the consummatory behavior may become active).

Learning a new appetitive behavior is accomplished by adding this Behavior to an existing Behavior Group. For example, consider a dog learning that sitting when her owner says "Sit" will often lead to a reward. In this case, the "sitting" Behavior together with a Releasing Mechanism that fires on the verbal command "Sit" are both copied into the Behavior Group associated with feeding, and thus performing this trick becomes one more strategy that the dog can use for getting food.

3.6 Inhibition and Level of Interest

Any creature has only limited resources to apply to satisfying its needs (e.g. it can only walk in one direction at a time), and thus there needs to be some mechanism to arbitrate among the competing Behaviors within a Behavior Group, since they cannot all be in charge at once. Moreover, once a creature is committed to satisfying a particular goal, it makes sense for it to continue pursuing that goal unless something significantly more important comes along.

We follow Minsky [Minsky86] and Ludlow [Ludlow76,80] in relying on a model of mutual inhibition both to arbitrate among competing behaviors in a Behavior Group and to provide control for behavioral persistence. Mutual inhibition usually (see [Blumberg94,Blumberg95]) leads to what Minsky terms the "avalanche effect," in which even weakly superior Behaviors can quickly come to dominate all the others. Ludlow provides two further important insights. The first is to recognize that by adjusting the inhibitory gains, one can vary the relative persistence of one Behavior versus the others. In general, the greater the gain, the more persistent the winning Behavior will be. The second insight is to associate a Level of Interest with every Behavior, indicating how "interested" the creature is in pursuing this Behavior right now. When the Behavior is active, its Level of Interest decreases, whether or not the Behavior achieved its goal. This in turn reduces the value of the Behavior regardless of its intrinsic value (i.e. its value before inhibition and Level of Interest were taken into account). Eventually, this will allow other Behaviors to become active. Level of Interest variables thus help the creature to avoid situations in which it obsesses on a single unattainable goal, performing the same currently ineffective Behavior over and over without result, and ignores less important but achievable goals.

Level of Interest also has a clear meaning in the context of an appetitive Behavior: it controls how long the creature should persist in the Behavior before giving up. For example, if the creature has learned from experience that a reward usually follows within X ticks of commencing some Behavior, then it should base the amount of time it will pursue this Behavior on this knowledge (i.e., it should accordingly adjust the amount by with the Level of Interest is decreased after each performance of the Behavior).

4 Integration of Learning

As indicated in Section 1, we want Silas to be able to learn: new contexts for goal-satisfying with existing behaviors; new releasing mechanisms for consummatory behaviors; and new releasing mechanisms with greater predictive

power that previously-used releasing mechanisms. The first two of these items are thinly veiled descriptions of what is usually referred to as operant or instrumental conditioning. In particular, learning new contexts for existing behaviors is essentially what animal training is all about (what Lorenz [Lorenz73] refers to as "conditioned action" or "conditioned appetitive behavior,". That is, the animal learns an association between an behavior and an outcome, and also learns the stimulus configuration upon which this association is contingent. The third type of learning corresponds more closely to "classical conditioning" in that there is already a built-in association between a stimulus and a response (i.e. a Releasing Mechanism and a Behavior), and what is learned are one or more novel Releasing Mechanisms that have an incremental predictive value (e.g. can predict an outcome further in the future). In the first two cases, a change in an underlying motivational variable acts as the reinforcement or feedback signal. By contrast, in the third case feedback is provided by an existing Releasing Mechanism. In the rest of this section we outline how these types of associative learning can be integrated into the action selection architecture described in Section 3.

4.1 Conceptual Overview of Learning Process

Fundamentally, the animal is trying to learn one of a number of potential associations:

- (S→O): Context S leads to outcome O, regardless of behavior
- (A→O): Behavior A leads to outcome O, regardless of context
- (S→(A→O)): Behavior A leads to outcome O, in context S

Here, outcomes mean a significant change in a motivational variable. Given these possible associations in the world, what should the learning system pay attention to and make associations with when a particular outcome occurs?

Temporal and spatial contiguity answers the first part of this question. That is, when a motivational variable changes, the learning system looks at Silas's recent behaviors and at the recently changed features of (nearby) objects of interest, to see if any of these may be associated with the outcome. Essentially, Silas says "let's check short term memory and start watching the last *N* behaviors and objects of interest to see if they continue to be associated with changes in this motivational variable from now on."

We use Sutton and Barto's temporal difference model of Pavlovian conditioning [Sutton90] to actually learn the prevailing associations. This model effectively forces alternative "explanations" (i.e., behaviors or object features) to compete to "explain" (i.e., make the best prediction of) the time-discounted changes in the value of the reinforcement variable (i.e., the outcome). Sutton and Barto's model is attractive because it is simple, yet it explains most of the results of Pavlovian conditioning experiments. It results in Silas learning a value for a particular behavior or stimulus that reflects its time-discounted association with changes in the value of some motivational variable.

In the first two types of associations, the change in the value of the underlying motivational variable (i.e., the outcome) acts as the reinforcement variable that drives learning, making simple first-order associations. In the third case (S→(A→O)), the learned value of A (in producing outcome O) acts as the reinforcement variable for learning O's association with S. This allows Silas to learn second-order associations.

The fundamental insight that allows us to integrate this learning algorithm into our existing action selection algorithm is that the "value" to be learned for a behavior or stimulus is represented by the MaxValue associated with the Releasing Mechanism sensitive to that behavior or stimulus. In order to treat behaviors as learnable "cues" in the same way as external stimuli, we use PRMs (Proprioceptive Releasing Mechanism) to signal when a given behavior is active. In all cases, then, we are learning the appropriate MaxValue associated with a given Releasing Mechanism.

When the learned MaxValues of the relevant Releasing Mechanisms get above a threshold, this means that the association is "worth remembering," and should be made a permanent part of the Behavior network. For example, in the case of a (S→(A→O)) association worth remembering, this means that in context S, signalled by one or more Releasing Mechanisms, performance of behavior A has reliably led to outcome O, which corresponds to the performance of a consummatory behavior that reduces the value of the underlying motivational variable. More concretely, in the case of sitting (A) on command (S) for food (O), the sitting behavior and the set of releasing mechanisms (e.g., an outstretched hand) that indicate it's time for Silas to do his trick will be added to the Behavior Group that is principally concerned with feeding. In addition, the Releasing Mechanisms that signal S will be added to the top-level behavior that owns the Feeding Behavior Group. This last step is necessary because S, the sitting-trick context, is now a relevant context for engaging in Feeding.

Because in this example (S→ (A→O)) represents a new appetitive strategy, it is also important to learn the appropriate time-course for the actions involved. That is, Silas should learn how long he should engage in the particular behavior of sitting, waiting for a cookie, before giving up. This information is computed as a by-product of the learning process and is used to change the Level Of Interest parameter of the behavior over time (see Section 4.2.6).

4.2 Extensions to Support Learning

To add learning to the action selection mechanisms presented in Section 3, we have developed several new structures that we now describe in turn.

4.2.1 Short-Term Memory

The Behavior System maintains FIFO memories that are essential for learning. One memory keeps track of the last *N* unique behaviors that have been active, and another keeps track of the last *N* objects of interest (currently we use *N* = 10). The active behavior on a time-tick is the leaf behavior which ultimately won during action selection. The object of interest is the pronome associated with that behavior. Note that in both cases the memory is based on unique instances and not time. This idea is taken from Killeen: "Decay of short-term memory is very slow in an undisturbed environ-

ment and dismayingly fast when other events are interpolated...Such distractors do not so much subvert attention while time elapses but rather by entering memory they move time along" [Killeen94].

Foner and Maes [Foner94] also emphasize the importance of contiguity in learning. However, because short-term memory in our system is based on unique behaviors or objects of interest, our focus of attention is not strictly equivalent to temporal or spatial contiguity. In addition, we will probably move to a system in which each motivational variable has its own memory-list of potentially significant objects or events, again stripped of their temporal information. This would then make it possible to model long-time-span learning phenomena such as taste aversion.

4.2.2 Reinforcement Variables

A Reinforcement Variable is either an Internal Variable (i.e. a motivational variable), or a Releasing Mechanism (either an ERM or a PRM) whose value is used as a feedback signal for learning. For example, hunger would be a Reinforcement Variable if the animal wished to learn new appetitive strategies to reduce its level of hunger. Alternatively, the PRM associated with the "Begging" behavior would be the Reinforcement Variable associated with learning the context in which Begging was a successful appetitive behavior for reducing hunger.

4.2.3 Discovery Groups and the Learning Equation

A Discovery Group is a simply a collection of Releasing Mechanisms (either PRMs or ERMs) for which the animal wants to discover the appropriate association with a Reinforcement Variable. The Releasing Mechanisms in a Discovery Group effectively compete to "explain" the value of the Reinforcement Variable, and as they learn their association with the Reinforcement Variable, they change their maximum value accordingly. Thus, a Releasing Mechanism that has no apparent association with the state of a Reinforcement Variable will wind up with a maximum value close to zero. By contrast, a Releasing Mechanism which has a strong association will wind up with a maximum value that approaches the time-discounted value of the Reinforcement Variable.

We now turn to the question of how Discovery Groups are formed, that is, how creatures "decide" what to watch as potential sources of important associations. The first point to be made is that it is there is no "centralized" learning mechanism. Rather, each motivational variable is responsible for learning its own set of appetitive behaviors and their respective contexts using the mechanisms described in the previous sections. This means that each motivational variable has its own collection of Discovery Groups for which it is the Reinforcement Variable. Thus, exogenous changes (i.e. changes directly due to the performance of a behavior) in the motivational variable will serve as the feedback for the learning process in these Discovery Groups. This process is summarized in Figure 5.

The second point is that significant changes in the level of the motivational variable initiate the process of identifying candidates (i.e. ERMs or PRMs) to add to the Discovery

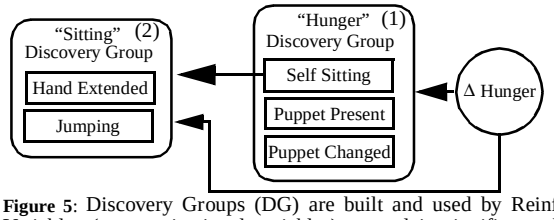

Figure 5: Discovery Groups (DG) are built and used by Reinforcement Variables (e.g. motivational variables) to explain significant changes in their value. For example, when hunger undergoes a significant change, it looks at short-term memory for the most recent behaviors performed and objects of interest encountered and adds the appropriate RMs (e.g. "Self-Sitting") to its primary DG (1). Within a DG, the Barto-Sutton TD learning algorithm is used to discover the appropriate MaxValue for each RM. In order to learn the appropriate context for a behavior which appears to be important, a subsidiary DiscoveryGroup (2) may be constructed which has both the motivational variable and the proprioceptive RM for that behavior as its reinforcement variables. It will be populated with RMs which are sensitive to those features (e.g. "Hand-Extended") of the recently encountered objects of interest which appear to change shortly before the motivational variable undergoes its change.

Groups. The fundamental assumption is one of contiguity: i.e., that one or more of the creature's recent behaviors and/or one or more of the features associated with the creature's recent objects of interest may have a causal or predictive association with the change in the motivational variable. In the current system, this means choosing candidates from the contents of short-term memory.

Discovery Groups and Reinforcement Variables are intended to be an abstraction of the temporal difference associative models proposed by researchers such as Klopf [Klopf93] or Sutton and Barto [Sutton90] to explain the results of classical conditioning. In their context, the Reinforcement Variable corresponds to the Unconditional Stimulus (UCS), and the Discovery Group corresponds to the collection of Conditional Stimuli (CS) that are competing to "explain" the value of the UCS.

We follow the Sutton and Barto TD learning model in this paper. Thus, within a Discovery Group we use the following equation to learn the association between a Reinforcement Variable and one or more Releasing Mechanisms:

$$\Delta V_{it} = \beta\left(\left[\lambda_t + \gamma\left[\sum_i V_{j(t-1)} \cdot A_{jt}\right]\right] - \left[\sum_i V_{j(t-1)} \cdot A_{j(t-1)}\right]\right)\left[\alpha \bar{A}_{it}\right]$$

Where:

V_{it} = MaxValue of RM i at time t

λ_t = Value of reinforcement variable at time t

$\bar{A}_{it}$ = Trace for RM i at time t

A_{it} = 1 if RM i passed find/filter and weight > .90, 0 otherwise

β = Learning Rate

γ = Temporal Discount Rate

α = Trace Rate

δ = Trace Decay Rate

Where the trace is defined as:

$$\bar{A}_{it} = \bar{A}_{i(t-1)} + \delta(A_{i(t-1)} - \bar{A}_{i(t-1)})$$

The first equation specifies how the MaxValue associated with a given RM in the Discovery Group will change in response to the value of the Reinforcement Variable. The amount it will change is proportional to an error defined to be the feedback actually received plus the discounted prediction of future feedback by all the Releasing Mechanisms in the Discovery Group less the prediction of feedback by all the Releasing Mechanisms on the previous time step. In

other words it changes by an amount proportional to the amount of actual and predicted feedback unexplained by the set of Releasing Mechanisms in the Discovery Group. In effect, they are competing for value, where the value is the time-discounted value of feedback.

The second equation is used to specify a "stimulus" or memory trace for a given Releasing Mechanism. It basically does temporal averaging of the state of the Releasing Mechanism, where A is either 1 or 0 (i.e. active or inactive) and A bar ranges between 0 and 1. This term is used by Sutton and Barto to explain temporally discontinuous associations between a CS and a UCS (e.g. when the CS turns off prior to the start of the UCS).

4.2.4 A Variable Learning Rate

A variable learning rate is a necessary addition to the learning equation in order to explain two results found in animal learning. First, prior exposure to a CS without a correlated UCS delays learning a later association between the two. Second, it is well known that partial reinforcement, while lengthening training times, also lengthens the time to extinction. Neither of these effects would be predicted by the standard fixed-rate learning equation. Gallistel [Gallistel90] and Cheng [Cheng95] base their arguments against associative models in part on just these reasons.

To rescue associative learning, we propose that the learning rate should be proportional to some measure of the observed reliability of receiving, or not receiving, feedback. If the reliability is high -- e.g., a behavior is always rewarded, or always not rewarded -- then it makes sense to have a high learning rate, i.e. to pay attention to each observation. On the other hand, if the observed reliability is low -- e.g., a behavior is sometimes rewarded, and sometimes not -- then the learning rate should be low, reflecting the uncertainty associated with any one observation.

To implement this intuition, we base the learning rate on a moving average of what we call the "Reliability Contrast" which is defined to be:

$$RC = \frac{\#(\text{active \& feedback})}{\#(\text{active})} - \frac{\#(\text{active \& no feedback})}{\#(\text{active})}$$

In other words, the RC is the difference between the observed probability of receiving feedback at least once while the behavior is active and the observed probability of receiving no feedback while the behavior is active. The RC ranges from reliably-unreliable (-1.0)—e.g. you will reliably not receive feedback to reliably-reliable (1.0)—e.g you will reliably receive feedback. The equation for the learning rate is then:

$$\beta = \beta_{min} + \text{fabs}(RC)(1 - \beta_{min})$$

This variable learning rate allows us to accommodate one of the two weaknesses of the learning equation, namely the partial reinforcement effect. It will also accommodate the phenomena of rapid re-acquisition after extinction because in extinction the learning rate is high, and so it responds to the change rapidly. Both effects are shown in Figure 6. It does not, however, accommodate the phenomena of prior exposure reducing the learning rate. We also do not yet model the transition from a learned response to essentially a habit (i.e. when the learning rate decays to 0 and learning stops).

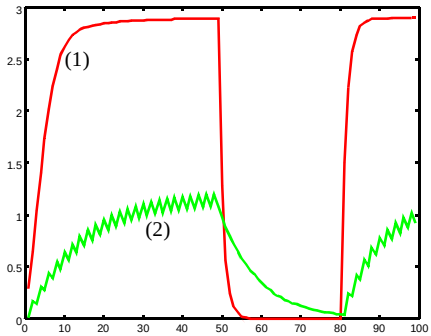

Figure 6: The graph shows the effect of a variable learning rate on the time-course in the associative strength of a stimulus under acquisition (t<50), extinction (50<t<80), and reacquisition (t>80) in 2 feedback ratio scenarios ((1) = rewarded each time, and (2) = rewarded every other time). The learning rate is calculated in each case on a moving average of the reliability contrast. Since the reliability contrast is lower in the case of the variable reward scenario (2), its learning rate is lower and thus its decay in extinction is more gradual.

4.2.5 Learning the Level of Interest

As mentioned earlier, in a world of uncertain rewards an animal needs to learn not just new appetitive behaviors, but also how long to engage in a given appetitive behavior before giving up and trying something else. We use a very simple approach to this problem: as part of the learning process, the Behavior incorporates its expected time to feedback and the associated variance. The Level of Interest for the Behavior is then held constant (and high) over the expected time to reward, and from that point on is allowed to decay linearly to zero over a period of time.

5 Implementation & Results

The algorithm described in the previous sections has been implemented as part of the Hamsterdam toolkit for building and controlling autonomous animated creatures. We have developed several creatures using this toolkit, including the ALIVE [Maes96] project's previously-mentioned virtual dog, Silas, a Puppet, and a Hamster [Blumberg95, Blumberg94]. Here we describe some of the initial results we have obtained for Silas and the hamster learning new positive and negative environmental associations.

Silas has 24 dog-specific Motor Skills and responds to 70 motor commands. In his current simplified form for testing learning, Silas initially has a Feeding Behavior Group which is empty except for a consummatory "eat" behavior--that is, this Behavior Group has no appetitive behaviors. Silas also has an exploratory Behavior Group that causes him to approach the user and randomly engage in a variety of different behaviors (e.g. sitting, begging, lying-down etc...).

We have used the learning algorithm described in the previous section to successfully train Silas to respond to gestures made by another autonomous creature, the Puppet (see Figure 1). The moment the Puppet detects that Silas is performing a desired behavior[2] (e.g. sitting), he performs the training gesture (e.g. hand-extended-down), and 10 ticks later (approximately 1/2 second), gives Silas a dog bone. Silas successfully learns that sitting when the Puppet's hand is extended is a useful appetitive strategy for getting food, and thus adds the "Sit" behavior and its associated RM for

"hand-extended" to its "Feeding" system.

As the next step in teaching Silas to sit, the Puppet (an unexperienced dog-trainer) changes his training gesture to be the combination of "hand extended" and "jumping." However, as a result of blocking, Silas attributes little value to the "jumping" component of the combined gesture, because it has no incremental predictive value beyond the "hand extended" component. But when the Puppet changes his training paradigm yet again and begins to jump several ticks before he extends his hand, Silas rapidly learns the new significance of "jumping". This is shown in Figure 7.

This experiment shows that the learning system described here can produce some of the phenomena associated with classical conditioning (not altogether surprising given our learning equation). Silas also learns the expected time to reward and adjusts the Level of Interest associated with "sitting" in this context accordingly, as discussed in Section 4.2.5.

In another experiment, we have a Hamster which is engaged in foraging in an open pen. The floor of the pen can be "electrified" and its color can change. The experiment is set up so that the floor-color will go from off to green to yellow to red. When the color is red, a shock is delivered for up to 40 ticks as long as the Hamster remains within a certain radius of the center of the pen. If the Hamster leaves the critical area, the shock is immediately turned off. The Hamster has a built-in response to shock: it will randomly choose to either "freeze" or to "run away." It also has a built-in association that the shock comes from the environment (i.e. when it is shocked the environment becomes its object of interest). The motivational variable is "relief from pain." Since the Hamster's speed is such that it can leave the pen in substantially less time than the shock's duration, running away should be the preferred action, and it should have an increasing tendency to run away as the floor-color progresses from green to red.

Over the course of 20 trials (i.e. 20 individual shocks), the Hamster learns that running away brings a quicker reduction in pain than does freezing, and so running away becomes the preferred option. In addition, it learns the association between the time course of the colors and being shocked and begins to flee when it senses the green light, or the yellow light at the latest. As a result, it successfully learns to avoid getting shocked.

We have also performed an experiment similar to that described by Montague [Montague94], in which our hamster learns to preferentially approach one food source over another based on the relative value of the alternative food sources. However, this needs to be combined with a motivationally-based exploration strategy so that the Hamster will periodically test the less-preferred source to see if its value has changed. These results, along with those for Silas's training, are still early and informal, but together they illustrate the power of the approach taken in this paper.

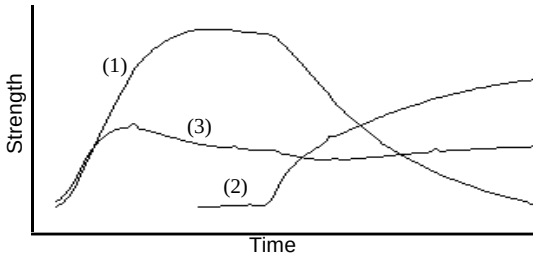

Figure 7: This shows the change in MaxValue for the "extendedRight" ERM (1), "jumping" ERM (2) and the "Sit" PRM (3) as a result of being associated with food during training. Initially, the Puppet uses a single gesture ("extendedRight"), but midway through training begins to simultaneously combine another gesture, "jumping". As a result of blocking, little associative strength is allocated to the "jumping" PRM (flat portion of 2) until the Puppet changes its training paradigm again and "jumping" precedes "extendedRight" by 3 ticks. As a result, "jumping" rapidly gains strength at the expense of "extendedRight".

6 Areas for future work

Using the learning algorithm presented here we plan on demonstrating further learning effects from the ethological and psychological literature. For example, we want to demonstrate different phenomena including taste aversion, outcome devaluation, habit formation, learned helplessness and superstitious behavior. From the ethological literature, we want to demonstrate socially-facilitated learning [Gould94, Shettleworth95] and integrate an exploration vs. exploitation strategy which is under the influence of motivational variables.We would also like to be able to demonstrate chaining to produce behavioral sequences (e.g. as in a circus routine).

Learning new motor skills (or combinations of motor skills) represents a major area of behavioral innovation which is not addressed here and thus represents a limitation in the current work. However, the architecture of the Motor System would make it straightforward to add this type of learning.

7 Conclusion

Our contribution in this paper is to show how certain types of learning may be integrated into the animat architecture previously proposed by Blumberg [Blumberg94, Blumberg95]. The underlying algorithm is Sutton and Barto's Temporal Difference model, but we show how it may be used and interpreted within the context of our ethologically inspired model to build and modify portions of the behavior network, and to set several of the fundamental parameters associated with our system.

It is not the intent of the authors to "push" one learning algorithm vs. another, but rather to stimulate thought on how best to integrate learning techniques such as TD learning into ethologically inspired multi-goal action selection architectures. This is done particularly with an eye towards building autonomous creatures such as "Silas" which must interact with, and learn from, users in natural ways.

2 Typically a dog trainer would issue the command "Sit" and then manipulate the dog into a sitting position [Rogerson92,Lorenz94]. Hence, the command precedes performance. To accomplish this in our system the Puppet considers the Dog to be performing a behavior the moment the behavior becomes active. However, the Dog (i.e. its PRMs) does not consider itself to be performing a behavior until both the behavior and its underlying motor skills are active. The effect is that from the Dog's perspective the command precedes its perception of its own performance.

We also hope that we have given the reader a sense of how ideas from ethology may be profitably incorporated into architectures for autonomous animated creatures.

References

Blumberg, B. (1994). Action-Selection in Hamsterdam: Lessons from Ethology. In: *From Animals To Animats, Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*, Cliff, D., P. Husbands, J.A. Meyer, and S.W. Wilson, eds. MIT Press, Cambridge Ma.

Blumberg, B. and T. Galyean (1995). Multi-level Direction of Autonomous Creatures for Real-Time Virtual Environments. In: *Proceedings of SIGGRAPH 95*.

Booker L. (1988). Classifier Systems that Learn Internal World Models. *Machine Learning Journal*, Volume 1, Number 2,3.

Brooks, R. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation* RA-2.

Cheng, P. and K.J. Holyoak (1995). Complex Adaptive Systems as Intuitive Statisticians: Causality, Contingency, and Prediction. In: *Comparative Approaches to Cognitive Science*, Roitblat, H.L. and J.A. Meyer, eds. MIT Press, Cambridge Ma.

Davey G. (1989). *Ecological Learning Theory.* Routledge Inc., London.

Dickinson, A. (1994). Instrumental Conditioning. In: *Animal Learning and Cognition*, Mackintosh, N.J. ed. Academic Press, San Diego.

Foner, L.N. and P. Maes (1994). Paying Attention to What's Important: Using Focus of Attention to Improve Unsupervised Learning. In: *From Animals To Animats, Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*, Cliff, D., P. Husbands, J.A. Meyer, and S.W. Wilson, eds. MIT Press, Cambridge Ma.

Gallistel, C.R. (1990). *The Organization of Learning.* MIT Press, Cambridge Ma.

Gallistel, C.R. (1994). Space and Time. In: *Animal Learning and Cognition.* Mackintosh, N.J. ed. Academic Press, San Diego.

Giszter, S. (1994). Reinforcement Tuning of Action Synthesis and Selection in a Virtual Frog. In: *From Animals To Animats, Proceedings of the Third International Conference on the Simulation of Adaptive Behavior*, Cliff, D., P. Husbands, J.A. Meyer, and S.W. Wilson, eds. MIT Press, Cambridge Ma.

Gould, J.L. and C.G. Gould (1994). *The Animal Mind.* Scientific American Library, New York.

Horswill, I. (1993). A Simple, Cheap, and Robust Visual Navigation System. In: *Second International Conference on the Simulation of Adaptive Behavior*. Honolulu, HI. MIT Press

Kaelbling, L. (1992). *Learning in Embedded Systems*, MIT Press, Cambridge Ma.

Killeen, P.R. (1994). Mathematical principles of reinforcement. *Behavioral and Brain Sciences*, 17,105-172.

Klopf, A.H., J.S. Morgan, and S.E. Weaver (1993). A Hierarchical Network of Control Systems that Learn: Modeling Nervous System Function During Classical and Instrumental Conditioning. *Adaptive Behavior*, Vol. 1, No. 3.

Lin, L.J., and T.M. Mitchell (1992). *Memory Approaches to Reinforcement Learning in Non-Markovian Domains*, CMU-CS-92-138, Dept. of Computer Science, Carnegie-Mellon University.

Lorenz, K. (1973). *Foundations of Ethology.* Springer-Verlag, New York.

Lorenz, K. (1994). *Man Meets Dog.* Kodansha America. New York.

Ludlow, A. (1976). *The Behavior of a Model Animal*. *Behavior*, Vol. 58.

Ludlow, A. (1980). The Evolution and Simulation of a Decision Maker. In: *Analysis of Motivational Processes*, Halliday F.T. &T. eds. Academic Press, London.

Maes, P. (1990a). Situated Agents Can Have Goals. *Journal of Robotics and Autonomous Systems* 6(1&2).

Maes P. & R. Brooks (1990b). Learning to Coordinate Behaviors. In: *Proceedings of AAAI-90.*

Maes, P. (1994). Modeling Adaptive Autonomous Agents. *Artificial Life*, Vol. 1, Numbers 1&2.

Maes, P., T. Darrell, B. Blumberg, and A. Pentland (1996). The ALIVE System: Wireless, Full-Body Interaction with Autonomous Agents, (to appear in the ACM Special Issue on Multimedia and Multisensory Virtual Worlds, spring 1996)

Mahadevan S. and J. Connell (1991). Automatic Programming of Behavior-Based Robots using Reinforcement Learning. In: *Proceedings of the Ninth National Conference on Artificial Intelligence*. MIT Press, Cambridge Ma.

McFarland, D. (1993). *Animal Behavior.* Longman Scientific and Technical, Harlow UK.

Miller, G.F., and P.M. Todd (1990). Exploring adaptive agency I: Theory and methods for simulating the evolution of learning. In: *Proceedings of the 1990 Connectionist Models Summer School*, Touretzky D.S., J.L. Elman, T.J. Sejnowski, and G.E. Hinton, eds. Morgan Kaufmann, San Mateo, Ca.

Minsky, M. (1988). *The Society of Mind*. Simon & Schuster, New York.

Montague, P.R., P. Dayan, and T.J. Sejnowski (1994). Foraging in an uncertain environment using predictive Hebbian learning. In: *Advances in Neural Information Processing 6*, Cowan, J.D., Tesauro, G, and Alspector, J. eds. Morgan Kaufmann, San Mateo, Ca.

Plotkin, H.C. (1993). *Darwin Machines and the Nature of Knowledge*. Harvard University Press, Cambridge.

Reynolds, C. W. (1987). Flocks, Herds, and Schools: A Distributed Behavioral Model. In: *Proceedings of SIGGRAPH 87*.

Rogerson, J. (1992). *Training Your Dog.* Howell Book House, London.

Shettleworth, S.J. (1994). Biological Approaches to the Study of Learning. In: *Animal Learning and Cognition*, Mackintosh, N.J. ed. Academic Press, San Diego.

Sutton, R., and A.G. Barto (1990). Time-Derivative Models of Pavlovian Reinforcement. In: *Learning and Computational Neuroscience: Foundations of Adaptive Networks.* Gabriel, M. and J. Moore, eds. MIT Press, Cambridge Ma.

Sutton R. (1991). Reinforcement Learning Architectures For Animats. In: *From Animals To Animats, Proceedings of the First International Conference on the Simulation of Adaptive Behavior*, Meyer, J.A. and S.W. Wilson, eds. MIT Press, Cambridge Ma.

Todd, P.M., and Miller, G.F. (1991). Exploring adaptive agency II: Simulating the evolution of associative learning. In: *From animals to animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, Meyer J.A., and S.W. Wilson, eds. MIT Press, Cambridge Ma.

Tu, Xiaoyuan and D. Terzopoulos (1994). Artificial Fishes: Physics, Locomotion, Perception, Behavior. In: *Proceedings of SIGGRAPH 94*.

Tyrrell T. (1993). *Computational Mechanisms for Action Selection.* Ph.D. Thesis, Centre for Cognitive Science, University of Edinburgh.

Watkins C. (1989). *Learning from Delayed Rewards*. Ph.D. Thesis, King's College, Cambridge.

Whitehead S.D. (1992). *Reinforcement Learning for the Adaptive Control of Perception and Action.* Technical Report 406, University of Rochester Computer Science Dept.

Woodhouse, B. (1982). *No bad dogs: the Woodhouse way.* Summit Books, New York.