Shaping Robot Behavior Using Principles from Instrumental Conditioning

Lisa M. Saksida[1,3]

Scott M. Raymond[2]

David S. Touretzky[2,3]

[1]Robotics Institute

[2]Computer Science Department

[3]Center for the Neural Basis of Cognition

Carnegie Mellon University

Pittsburgh, PA, USA 15213

saksida@ri.cmu.edu, sr4r@andrew.cmu.edu, dst@cs.cmu.edu

Address correspondence to Lisa M. Saksida.

Phone: 412-268-1417

Fax: 412-268-5571

Abstract

Shaping by successive approximations is an important animal training technique in which behavior is gradually adjusted in response to strategically timed reinforcements. We describe a computational model of this shaping process and its implementation on a mobile robot. Innate behaviors in our model are sequences of actions and enabling conditions, and shaping is a *behavior editing* process realized by multiple editing mechanisms. The model replicates some fundamental phenomena associated with instrumental learning in animals, and allows an RWI B21 robot to learn several distinct tasks derived from the same innate behavior.

Keywords: autonomous mobile robots, instrumental learning, operant conditioning, reinforcement learning, shaping.

1. Introduction

Service dogs trained to assist a disabled person will respond to over 60 verbal commands to, for example, turn on lights, open a refrigerator door, or retrieve a dropped object [9]. Chicks can be taught to play a toy piano (peck out a key sequence until a reinforcement is received at the end of the tune) [6], and rats have been conditioned to perform complex memory tasks which are analogs of human cognitive tests [8, 18]. These and many other complicated behaviors can be acquired as a result of the delivery of well-timed reinforcements from a human trainer. This training strategy exploits a type of learning found from invertebrates to humans, in which an animal comes to associate its actions with the subsequent consequences. In the animal learning literature, this is referred to as instrumental, or operant, conditioning. One specific animal training technique which exploits this type of learning is *shaping by successive approximations*, or just "shaping".

Mobile robot learning algorithms have yet to approach the sophistication and robustness of animal learning. While the idea of reinforcement for appropriate actions is commonplace in the machine learning literature [21], little of that work examines animal training methods. Several researchers have developed robot "shaping" systems [11, 28], and these methods are a significant contribution to the robot learning literature, but they have not addressed behavioral shaping in animals in any detailed way. Efforts toward understanding instrumental learning through computer simulations have, to this point, addressed only elementary phenomena such as encouraging or suppressing a single motor action [3, 30]. A serious attempt to model phenomena described in the instrumental learning literature will serve the dual purpose of improving the abilities of robot learners while at the same time yielding a fresh, computationally-oriented perspective on animal learning.

In the present paper we develop a computational theory of shaping that is at a level appropriate for generating mobile robot tasks. Using animal learning theory as a basis, we propose the idea of *behavior editing* as a potential mechanism underlying this learning process. Behavior editing is a method by which a pre-existing behavior, either innate or previously learned, can be changed in magnitude, shifted in direction, or otherwise manipulated to produce a new type of behavior. To demonstrate this, we have implemented our model on Amelia, an RWI B21 mobile robot. We provide results from training Amelia on several tasks, all of which are built from one innate behavioral routine, and discuss how these results reflect shaping in animals.

2. Associative Learning in Animals

2.1. Pavlovian and Instrumental Conditioning

Two basic types of associative learning — Pavlovian (or classical) conditioning and instrumental (or operant) conditioning — are well established in the animal learning literature. Although the line between these two types of learning is often blurred [7], they are considered to reflect different underlying processes [12].

Pavlovian conditioning results in the association of an arbitrary neutral stimulus called a conditioned stimulus (CS) with a second, unconditioned stimulus (US). The US is inherently pleasant or unpleasant, and thus provokes an innate, unconditioned response. The key factor in this type of learning is that if the CS reliably precedes the US, eventually the CS will elicit the same response as the US. For example, food naturally produces a salivation response in animals, thus it is a US. If an animal is consistently presented with a tone (a CS since it produces no innate response) ten seconds before it receives food, eventually it will salivate in response to the tone itself, before the food is presented. At this point the tone has become predictive of food, and the animal has learned a CS$\rightarrow$US association. This suggests that Pavlovian conditioning helps an animal to recognize structure in its environment, enabling it to make useful predictions about upcoming events based on its perceptions.

During Pavlovian conditioning, the animal learns associations between stimuli in the environment, but it learns nothing about the effects of its own actions. Since the behaviors which are generated are hardwired to occur in relation to the US, the animal depends on its genetic heritage for appropriate responses to stimuli. If the environment changes such that certain actions are now much better or much worse than before, an animal with only Pavlovian conditioning abilities would be incapable of adapting. This is where instrumental learning comes in. In an instrumental learning situation, the execution of an action, not just the occurrence of a stimulus, is required to produce a certain consequence. If that consequence is desirable, the action associated with it becomes more likely to occur. In other words, the learner constructs and maintains an association between an action and its outcome (A→O). The context in which this is learned also becomes incorporated into the association, such that the animal expects a given action to lead to that outcome only in the presence of specific stimuli, i.e., S→(A→O).

2.2. Shaping by Successive Approximations

Shaping by successive approximations is a top-down approach to behavior modification. It exploits innate or previously learned behaviors of the animal, which can be selected and modified through reinforcement to yield the specific behaviors desired by the trainer.

Consider an animal trainer teaching a dog to sit up and "beg". In order to produce this type of behavior, the trainer does not have to start from scratch, but instead can motivate the dog to emit an innate [16, 27] sequence of actions for pursuing an object (see Figure 1), and use appropriate reinforcement to mold the sequence into the begging behavior.

==

Figure 1 about here.

==

In order to get the dog to sit up on its haunches, the trainer dangles a piece of food above its head. The trainer also gives a verbal command ("beg"), which initially has no meaning to the animal but comes to be associated with the desired behavior. As soon as the dog is in the desired position (sitting on its haunches with its front paws off the ground), the trainer gives it praise or a food reinforcement, which interrupts the object-pursuit sequence with a desirable outcome (Figure 2).

==

Figure 2 about here.

==

After several trials, when the dog is reliably sitting up for food, the trainer initates a trial but holds the food several centimeters higher than previously.

On each subsequent trial, the trainer moves the food farther from the dog. As the distance to the food is increased, and the dog continues to receive reinforcement during the approach stage, the relevance of the sight of the food as a stimulus triggering the behavior decreases. At the same time, since the "beg" command has been consistently presented, the strength of the association between the command and the rewarded behavior increases (Figure 3).

==

Figure 3 about here.

==

Eventually the sight of the now distant food target becomes unnecessary, and the dog will perform the desired behavior in response to just the verbal cue. (Figure 4).

==

Figure 4 about here.

==

Once the dog is reliably sitting up in response to the "beg" command, the trainer can shape various parameters of the behavior. In order to do this, the trainer relies on natural variation in the magnitude of the animal's actions. In this case, if the trainer wants to teach the dog to hold the begging position for ten seconds, he or she would reinforce trials on which it begs for a relatively long time. If this is done consistently, eventually the average duration of the gesture will increase, and the experimenter can shift the criterion for reward higher. Eventually, when the average duration has been extended to around ten seconds, the trainer can begin reinforcing the dog only for actions that are very close to that duration, which produces a more consistent behavior.

Although learning to beg may at first appear to be a unitary phenomenon, a closer examination suggests that it is the result of a combination of several mechanisms acting together to modify extant actions and produce the observed behavioral changes. Shaping can be broken down into key components as follows:

1. The dog is learning a new behavior that can be triggered by the "beg" command. At the same

time, the original object-capture sequence will be emitted in response to interesting objects if no command is given. This reflects the fact that animals can learn many potentially overlapping behaviors which are triggered selectively depending on the situation.

2. The dog initially executes a behavior appropriate to attaining the goal of food. This is due to the fact that the dog has innately specified or previously learned an A→O association: execution of the object capture sequence will lead to a desirable outcome.

3. Initiation of the behavior sequence is triggered by the sight of the food target. This reflects the fact that stimuli can drive behavior. In the animal learning literature, this is known as *stimulus control*, and is thought to result from an S→(A→O) association.

4. When the reward is given before the dog completes the object capture sequence, the behavior is interrupted but with an overall positive outcome. The dog learns that successfully capturing the object is not necessary to earn a reward.

5. By the end of training, the sight of a food target is no longer necessary for the dog to emit the (now modified) object capture behavior. The procedure by which control of a visual stimulus is weakened by moving it further from the dog on successive trials is an instance of what is called *fading* in the animal training literature.

6. As the target is faded, the "beg" command becomes sufficient to trigger the desired behavior. This is an instance of transfer of stimulus control, and demonstrates that, though learned association, arbitrary stimuli can gain control over a behavior.

7. Once a brief "beg" behavior has been learned, the duration of the gesture is gradually increased by selectively rewarding longer instances. This is an example of adjusting the *topography* of a response.

Our model of animal learning utilizes several learning mechanisms to produce these different aspects of shaping. The fundamental concept is that these mechanisms operate on previously learned or innate behavior sequences by *editing* them in various ways, so that eventually they are transformed into the desired behavior.

3. A Simple Model of Animal and Robot Behavior

In order to develop our theory of behavior editing, we first introduce a model of animal behavior that can be implemented on a mobile robot. While admittedly simplistic, this model provides a concrete description of behavior which can then be subjected to our editing operations. More sophisticated behavior models should also be susceptible to editing.

We define a behavioral sequence as a collection of states, each with an associated action. Some states have abstract actions that can be instantiated in several ways by more specific states. For example, "move arm" can be realized by actions such as raising the arm, lowering the arm, waving, and so on. "Activation links" connect abstract states with their more concrete realizations.

The animal is always in some state, meaning that state is "active". The active state, if abstract, may have one of its realizations also active at the same time. If this realization is itself an abstract state, then it too may activate a realization, and so on. So at any given time there is a set of active states forming a chain of activation links from most to least abstract.

Action sequencing is accomplished by "transition links" between states at the same level of abstraction. A behavioral sequence is executed by following the chain of transition links from the start state to some goal state, performing the action associated with each state along the way. Transition links are weighted; the weights determine the likelihood that a given path will be selected from the choices currently available.

==

Figure 5 about here.

==

Figure 5 shows a behavior sequence for locating and manipulating objects. It can serve as a simplified, first-order description of animal prey-catching behavior. On the robot, this sequence results in a tendency to be attracted to objects. Once the robot sees something it recognizes, it approaches the object and tries to pick it up if graspable. Upon successful completion of this sequence the robot looks for other objects with which to interact.

4. Preconditions and Postconditions

States become active in particular contexts, i.e., certain conditions must hold. These are known as the preconditions for the state. Once activated, the state remains active until its postconditions are satisfied. Performing the action associated with the state (or one of its sub-states) should make the postconditions true. See [15] for a similar conception of states.

Preconditions are typically references to stimulus features such as the color of an object, while postconditions may reference either external relationships, such as the distance between the robot and an object, or internal variables such as the distance traveled by the robot since it entered the current state.

In our simplified model, each precondition is either a boolean quantity (e.g., whether or not a target has been detected) or a point in a continuous vector space defined by the stimulus modality. Boolean preconditions are satisfied (have match strength $M = 1.0$) when the feature value is 1 ("true"). Conditions on continuous-valued features are satisfied to the extent that the feature lies within a specified distance of some target value. We use a Gaussian classifier to judge the strength of match between a stimulus feature and the target value. Let μ_i be the target vector for the ith stimulus class, $x_i(t)$ be the value of that stimulus at time t, and σ_i^2 the variance in the distance of points from the target, i.e., the width of the tuning curve. Then the match strength for the ith stimulus is defined as

$$M_i(t) = \exp\left(\frac{- \parallel x_i(t) - \mu_i \parallel^2}{\sigma_i^2}\right) \tag{1}$$

If σ_i^2 is not too small, the model will generalize from a prototypical stimulus value to similar values, e.g., if trained to respond to an orange object, it will also respond to pink or yellow ones. However, the model can be taught to make arbitrarily fine stimulus discriminations by tightening the tuning curve.

Animals usually learn about certain types of stimuli more easily, or more quickly, than others [16]. In pigeons, for example, visual stimuli are more salient than auditory stimuli. In the model, each stimulus class has an innate salience, or *associability* (α_i), that determines its relative strength in influencing the animal's behavior. The salience of an individual stimulus instance also depends on its intensity, e.g., the brightness of a light, the loudness of a tone, or the proximity of a visually-tracked target [26].

Preconditions can be positive or negative, depending on the strength of association V_i between stimulus match and reward. For example, if pressing a lever produces a food reward except when a light is on, the light would develop a significant negative associative strength. The V_i terms are learned parameters and normally start at zero, but for innate behaviors with built-in conditions the V_i values may start at any value between -1 and $+1$.

==

Figure 6 about here.

==

Reinforcement of an action contributes to bringing that behavior under the control of the stimuli that were present at that time. That is, the probability of re-entering the state responsible for that action will be higher in the presence of the stimuli and lower in its absence [32]. In our model, stimuli gain control over behavior based on their relative match strengths, saliences, and associative strengths. We define the amount of control possessed by a stimulus feature to be

$$C_i(t) = M_i(t) \cdot \alpha_i(t) \cdot V_i \tag{2}$$

Figure 6 shows the target-color precondition, with its associated parameters and computed level of control.

The likelihood of entering a state is determined by the sum of the control levels of all preconditions of that state. If the stimuli which form preconditions for a state are present and have an aggregate degree of control that is above threshold, then the state will be entered. Similarly, the aggregate control level of all postconditions of the state must be above threshold for the state to be exited. Consider the *approach-object* state, when the robot is 100 cm from the target. Figure 7 shows that to enter this state, the target-present precondition must be satisfied. Since target-present is a built-in precondition, it has an initially high associative strength V_i. When a target is present there will be a high match value M_i and salience α_i; thus, the precondition will be satisfied and the robot will enter the approach-object state. It will remain in that state until the aggregate control level of the three postconditions (target-within-reach, target-distance, and distance-moved) exceeds the postcondition threshold. Each of these postconditions contributes to the successful completion of the approach state. Target-within-reach is satisfied when the robot is within reach of the target, target-distance records the distance the target is from the robot upon completion of the approach (satisfaction is contingent on the value being equal to the within-reach distance initially, but can change with reward), and distance-moved records the total distance the robot has moved during the approach stage of the trial.

==

Figure 7 about here.

==

Postcondition control changes over time as follows. Assuming that associability α_i remains constant (i.e., stimulus intensity does not change), control is dependent on M_i and V_i. For two of the postconditions, target-distance and distance-moved, the variance σ_i^2 associated with the expected value μ_i starts high, since it is not known in advance whether any particular target distance or movement distance will lead to reward. As a result, M_i for these postconditions will be high for any distance between the target and the robot. However, since the association between these postconditions and reward in the given situation is unknown, V_i will be low, and these two postconditions will not have a great deal of control. In contrast, since the target-within-reach postcondition is critical for moving out of the innate approach-object state, its initial V_i is high. Also, the target-within-reach condition has a specific requirement that is known in advance — the robot must be within 5 cm of the target in order for it to be within reach — so σ_i is tight and M_i will be low when the robot is far from the target. Once the approach-object state is active, however, the robot moves forward and the distance to the target decreases, so M_i for target-within-reach increases. Eventually, the robot comes close enough to the target that the total postcondition control exceeds the threshold, and the robot can exit the approach-object state (see Figure 8).

==

Figure 8 about here.

==

5. Learning Mechanisms (Behavior Editing)

Several different mechanisms can edit a behavior sequence. First, the sequence may be truncated by introducing a reward that interrupts the normal flow of state transitions. Second, the expected value and width of the tuning curve for a precondition can be adjusted, which alters the way the model generalizes or discriminates among stimuli. Third, the degree to which particular stimuli have control over a behavior can be manipulated by changing the associability α_i and associative strength V_i. Finally, the topography of an action can be altered by adjusting its parameters. These editing mechanisms are discussed in detail below.

5.1. Modifying State Transitions

On the first training trial, the robot executes an innate behavior sequence. It scans the room for targets and if it sees one, approaches it and possibly picks it up. If a reward is received during the course of this trial, a copy is made of the portion of the behavior sequence that was executed, and the various changes that can be made by the editing mechanisms are made to the copy, not to the original sequence. On later trials, if the stimuli in the environment sufficiently match the preconditions for the first state in the copied sequence, then that sequence is initiated, and any further changes made through reward are made to the copy. If the stimuli do not match the preconditions for any of the copied behavior sequences, the innate sequence is chosen. The function of these copies is to preserve the innate behavior and allow any number of variants to be learned, each of which is initiated only in the appropriate context.

When our model receives a reinforcement, it builds an expectation that reward will be obtained the next time it is in the current state, and it maintains a value for each state reflecting the strength of this expectation. This value is increased for the current state whenever a reward is received, and decreased when the reward expectation is above zero but a reward is not received. This expectation value can be seen as an A$\rightarrow$O association.

When the model receives a reinforcement, it moves into a "collect reward" state, and the reward expectation parameter for the previous state is increased. In addition, the strengths of all of the transition and activation links followed since the start of the trial are increased. But if a violation of the A$\rightarrow$O association occurs (the model does not receive a reward for executing an action when it was expecting one), the strengths of all transitions taken during the trial are decreased. In addition, the value of the reward expectation is reduced, thereby decreasing the strength of belief that the action leads to reward. With repeated non-reinforcing trials, eventually the transition strength for that action decreases to a point where the action ceases to occur any more frequently than other actions. In the animal learning literature, this process is known as extinction.

5.2. Acquisition of Stimulus Control: Addition and Modification of Preconditions

Reinforcement of an action contributes to bringing that behavior under the control of the stimuli that were present at that time. That is, the probability of reentering the state responsible for that action will be higher in the presence of the stimuli and lower in its absence [32]. The stimuli are then known as controlling, or discriminative, stimuli. Discriminative stimuli are important because they allow the animal to restrict its actions to only those situations in which they will produce reinforcement. Animal trainers take advantage of this by establishing particular gestures or verbal cues, such as the "beg" command in the dog training example, to place behaviors under stimulus control so that the animal performs on command. Undesirable behaviors can sometimes be eliminated by bringing them under stimulus control and then not giving the command any more [29].

As mentioned previously, in our model stimulus control is the product of associability (α_i), match strength (M_i), and associative strength (V_i). Since associability is dependent on stimulus intensity (i.e., increases for intense stimuli and decreases for fainter stimuli) and match strength reflects the similarity of the current stimulus to the expected one, both of these variables depend on current stimulus conditions. Therefore they can vary from trial to trial.

The associative strength V_i of a stimulus normally starts at zero and increases as the precondition is paired with reinforcement. It is adjusted in our model using the Rescorla-Wagner learning rule [31] (mathematically equivalent to the LMS or Widrow-Hoff learning rule [33]) when either a reward is received or an expected reward is not received. The learning rule for adjusting V_i is:

$$\Delta V_i(t) = \alpha_i(t) M_i(t) \left[\lambda(i, \text{reward}) - \sum_j \alpha_j(t) V_j(t) M_j(t) \right] \quad (3)$$

where $\alpha_i(t)$ is the associability of stimulus i, M_i is the match strength of the precondition, and λ is the target strength (amount of association supported by the reinforcer.) A λ value of 1 was used for reward, and 0 for no reward.

When a stimulus present during a reward is added as a new precondition, it has an initial associative strength of 0. If perceptions that closely match this stimulus are repeatedly accompanied by reinforcement, the associative strength gradually increases. Perceptions that don't match the stimulus very well don't appreciably affect the strength. Finally, if a particular stimulus is consistently associated with non-reward when reward is expected (e.g., a previously rewarded action sequence is no longer being rewarded), the stimulus will develop a negative associative strength, thus leading to a suppression of the sequence in the presence of that stimulus.

Use of the Rescorla-Wagner rule gives us an estimate of a stimulus' correlation with reward. It also allows us to replicate certain well-known phenomena in the classical conditioning literature, such as blocking and negative conditioning.[1] Our implementation of the rule differs from the original version described by Rescorla and Wagner [31] in two ways: first, we have a dynamic α (which is described in the next section) and second, we use a continuous value for stimulus match strength instead of a boolean indicator of stimulus presence.

5.3. Transfer of Stimulus Control: Fading

When a dog is learning to beg, it initially sits up on its haunches because it is trying to to reach the food being dangled above its head. It is performing an object-capture behavior. With training, the dog learns that sitting up in response to the "beg" command is rewarded. It no longer requires a visual target, and will execute the behavior on cue even if no target is present.

[1] In blocking, when an animal is first trained on stimulus A paired with reward, and then on stimuli A and B together with reward, it shows little or no response when tested on B alone. In negative conditioning, training on stimulus A paired with reward and then A plus B and no reward produces a negative associative strength for stimulus B.

To reach this state, the trainer slowly fades the target by holding it farther from the dog on each trial, until it can be eliminated altogether. During this time, control of the behavior is being transferred from the visual stimulus to the verbal command.

We model the fading process as follows. The stimulus intensity of a visual target is inversely related to its distance. As a target is moved further away, its intensity diminishes and since the associability of the stimulus is directly dependent on intensity, α_i decreases as well. Because we are essentially decreasing the learning rate for the target, even though M_i is still high, the target gains less associative strength as it is moved further away. The verbal command remains at full associability, however, so it is able to gain much more associative strength. In addition, since control is the product of α_i, M_i, and V_i, the lowered α_i value of the target results in a smaller degree of control for the target.

What happens if the controlling stimulus is faded too quickly or removed entirely? If the trainer pulls the target back from the dog before the "beg" command has acquired sufficient associative strength of its own, the dog will not respond appropriately. In the model, the total control possessed by the stimuli in the current context will be below threshold, because α_i for the target is small while V_j has not increased enough to give the verbal cue sufficient control of the behavior.

5.4. Stimulus Discrimination and Generalization

Implicit in the development of stimulus control is another type of learning. In order for the model to acquire knowledge about a context, it must be able to accurately discriminate the stimuli associated with that context from other stimuli encountered in the environment. On the other hand, it must not be too picky in distinguishing among stimuli; otherwise it may fail to recognize an appropriate context due to an insignificant variation in some stimulus quality.

Stimulus discrimination and generalization are controlled by the values of μ, the expected feature value, and σ^2, the variance or tuning curve width. Two processes adjust these parameters. Every time a reward is received or an expected reward is not received, the mean expected value and the variance of each active precondition are updated as follows:

$$
\begin{aligned}
N_t &= \sum_{i=1}^{t} \gamma^{i-1} && (4)\\
\mu_t &= \frac{1}{N_t}\sum_{i=1}^{t} \gamma^{i-1} x_i && (5)\\
S_t &= \| x_t - \mu_t \|^2 + \gamma \psi_{t-1} && (6)\\
\psi_t &= \begin{cases} 0 & \text{for } t = 0 \\ \eta S_t & \text{if reward expected but not received} \\ max(S_t, N_t, \sigma^2_{min}) & \text{if } S_{t-1} \geq N_{t-1}\sigma^2_{min} \\ S_t & \text{otherwise} \end{cases} && (7)\\
\sigma^2_t &= \frac{1}{N_t}\psi_t && (8)
\end{aligned}
$$

where γ is a discount factor slightly below 1, t is the number of trials, N_t is the time-discounted number of trials, S_t is the sum of discounted σ^2 values, and η is a variance shrinkage term (explained below).

When a reward is received, in order to maintain the possibility of generalization, σ^2_t is not permitted to shrink below a certain minimum value σ^2_{min} unless reward expectations have recently been violated. So even if the model is trained exclusively on, say, orange stimuli, the value of σ^2 should still be large enough to permit generalization to pink or yellow stimuli, though probably not dark blue ones. If a variety of different stimulus values are associated with reinforcement, then the variance σ^2 will increase to well above σ^2_{min} to reflect this.

When an expected reward is not received, the model may have failed to distinguish a desirable stimulus from a less desirable one. To correct this, the model narrows the tuning curve slightly by reducing S_t by a certain proportion η. This contraction is not subject to the minimum variance constraint. Hence, the model can be taught to sharpen its tuning curve to distinguish pink from orange stimuli if the reinforcement contingencies require this.

5.5. Shaping Action Topography

Yet another form of learning that occurs within a shaping procedure is the adjustment of individual actions. Since motor behavior is inherently variable, each instance of an action will tend to be slightly different. One can change the shape of the action by rewarding only those instances that come close to the desired form. In the above example, once the dog is sitting up in response to the "beg" command, the trainer can begin to modify the parameters of the gesture. If only the trials in which the dog begs for a relatively long length of time are reinforced, then the mean duration of the behavior will become longer.

A hill climbing mechanism is used to shape the motor action associated with a state. Each state can have postconditions which reference parameters of the motor action and are updated by adjusting μ and σ values in a similar manner to the preconditions discussed in the previous section. For example, the approach-object action has three postconditions: the robot must be within reach of the target, it must be within a certain (learnable) distance of the target, and it must have moved forward a certain distance. These postconditions must all be satisfied to an extent related to their strength of association in order for the robot to leave the state. When a target is present, the robot can approach the target to satisfy all of the postconditions. When the target has been deleted through fading, however, distance-moved becomes the only relevant postcondition. In this case, the learner selects parameter values for the move-forward motor action based on the distance-moved distribution described by μ and σ.

To shape the robot to move forward a specific distance, we first train it to move forward by triggering the object-capture sequence with a target. Next, we fade the target until a verbal cue is the only controlling stimulus. Finally, we only reinforce the robot for movements that are closer than average to the desired distance. This shifts μ in the direction we want to go and keeps σ broad enough that we are likely to see new actions that are closer to what we desire. Once the target value for distance traveled has been reached, we can adopt more stringent performance criteria, which tightens σ and encourages uniformity of behavior.

6. Robot Implementation

Amelia is an RWI B21 mobile robot (see Figure 9) with a color camera on a pan-tilt head and a three degree-of-freedom arm with gripper. Computing power is provided by three on-board Pentium processors; a 1 Mbps radio modem links Amelia to a network of high-end workstations that can contribute additional processing cycles. For our experiments, we ran the learning program in Allegro Common Lisp on a Sparc 5 and used TCA [34] to communicate with the robot.

==

Figure 9 about here.

==

In order to allow the robot to operate in real time, we chose a fast and simple approach to robot vision. Objects were assumed to be monochromatic. We defined canonical RGB values for each object type and used a gaussian classifier to label each pixel in the image. A connected components algorithm found contiguous regions bearing the same label. Then, a hand-tuned post-processor rejected regions that were too small to be objects, and merged regions with identical labels that were close enough together that they probably belonged to the same object. This provided some robustness against breaks in the image cause by shadows or small occlusions.

The human trainer wore a bright orange jacket which presented an easily-tracked monochromatic target. A second class of objects known to the robot was plastic dog toys of various colors; pink and green toys were used in the experiments described below. These toys had a symmetric four-armed shape that allowed them to be picked up in any orientation, so precise positioning of the arm was not required. The final object class was blue plastic recycling bins, which served as toyboxes.

Due to the camera's relatively narrow field of view, the robot was programmed to pan the room every 15 seconds, taking six snapshots arranged in a 2×3 array. The top row of images gave a good view of people in the room; the bottom row allowed the robot to see objects on the floor. The six snapshots were merged into one large image before processing. In between pan events the robot processed images with the camera pointing straight ahead at a rate of several frames per second.

The vision module returned a list of the objects it found, along with their actual colors and positions in the image. This list was converted into a list of simuli, triggering actions appropriate to the objects currently in view and their locations. Stimulus values were also referenced by the learning rule for preconditions discussed above to adjust the μ and σ parameters.

The human trainer delivered reinforcements to the robot via a Logitech three-button radio trackball. Standing anywhere in the vicinity of the robot, the trainer would press the right mouse button each time a desired response occured. The button press was picked up by a receiver plugged into a serial port on one of the robot's Pentium computers and relayed to the learning program on the Sparc. The robot did not receive feedback on every timestep or for each action performed; it only received a reinforcement when an appropriate action had occured.

7. Results

Experiments 1, 2 and 4 below were run on the actual robot using an earlier version of the model than is described here. Experiments 3, 5, and 6 were run in the current version of the model to demonstrate recent enhancements, such as stimulus fading and action topography shaping. These latter experiments were run in simulation, using visual inputs previously recorded from the robot's actual vision system.

"Verbal" commands were given to the robot at the beginning of trials. With a real animal these would be spoken words, but since we do not have a speech recognition system on the robot we used keyboard input instead. These commands, which were learned as new preconditions, allow us to teach the robot to perform a particular action only in the context of a verbal cue, as an animal trainer would do.

Prior to starting the experiments, a baseline measure of performance was generated by allowing the robot to execute its built-in action sequence. When the robot was shown the pink toy, for example, it approached it, lowered its arm, and picked it up.

7.1. Learning the Consequences of Actions by Truncating the Sequence

The purpose of this experiment was to demonstrate that new behaviors can be derived from an existing action sequence by changing the outcome of actions through reinforcement. The robot was taught to move forward in the presence of the pink toy without picking it up. First, the pink toy was introduced and the "move forward" command (initially meaningless to the robot) was given. The visual stimulus triggered the innate behavior sequence described above. As soon as the robot had moved 20 cm toward the toy, it was interrupted by a reinforcement from the radio trackball, which caused it to stop approaching the toy as it entered the "collect reward" state. After 5 training trials, we tested whether the robot's behavior had been modified by showing it the toy and giving it the command, but not interrupting it with a reward. If nothing had been learned, we would expect the robot to approach and pick up the toy. During this trial, however, the robot completed the approach to the target and stopped. This simple example illustrates that new behaviors can be acquired by interrupting an innate sequence with reward. This is similar to the first stage of training the dog to beg.

7.2. Follow the Trainer

The robot does not have to be trained in discrete trials; it can learn in a continuous exploration mode. In this experiment, the trainer wore the bright orange jacket and the general exploratory sequence was run on the robot continuously. Low-level collision avoidance software, which uses sonar to detect and avoid obstacles, was run along with the learning program. The trainer carried the trackball around and rewarded the robot whenever it approached her.

At the start of training, the robot wandered around the room and approached any object (within its repertoire of visible objects) that it saw. After several rewarded approaches to the trainer, the robot began to preferentially track and follow the trainer. By the end of 10 training trials, the robot was ignoring other objects in the room and following the trainer exclusively.

7.3. Stimulus Discrimination

In this experiment the program was taught to respond to green toys but not to pink ones. No commands were used: the program was simply given 30 trials of alternating green and pink inputs. It was rewarded for moving 20 cm toward a green toy but was not rewarded for approaching a pink toy. Since no verbal commands were given to distinguish the two situations, learning the discrimination was completely dependent on tuning the μ and σ values for the target-color precondition. By the end of training, match strength for the pink toy had decreased to a negligible value, whereas match strength for the green toy was close to 1.0. This led to different behavior on green and pink trials: on green trials the program approached the toy and expected reinforcement after moving forward 20 cm, whereas on pink trials the robot simply executed the default behavior sequence, since the learned behavior's preconditions were not adequately satisfied.

7.4. Recycling and Playing Fetch

In this experiment, the robot was taught to recycle green toys. First it was given 10 trials in which it was initially presented with the green toy and then the recycling bin. At the start of each trial, a "recycle green" command was given. As in the above stimulus discrimination experiment, the robot was rewarded after it picked up the green toy. It was not rewarded for picking up other toys. Next, the robot was given a "take to bin" command and was rewarded if it approached the recycling bin and dropped the toy in it.

During the first phase of training, the robot quickly learned the green toy recycling task. By the end of 10 training trials it was performing reliably, always picking up the green toys and taking them to the recycling bin even when other toys were available.

A variation on this theme involves teaching the robot to play fetch. Ten trials were given in which the robot was first presented with a green toy, followed by the trainer coming into view. At the start of the trial, the robot was given a "fetch" command and was rewarded whenever it retrieved the toy. Next, the robot was told "bring it here", and was rewarded when it had brought the toy to the trainer. No distractor targets were present (i.e., on each trial only the green toy or only the trainer could be seen), which made learning occur very quickly. Within 10 trials the trainer could throw the green toy anywhere in the room and the robot would search until it found it, then approach and pick up the toy, scan the room for the trainer, and finally approach the trainer and drop the toy at her feet.

7.5. Transfer of Stimulus Control (Fading)

This experiment demonstrates how the pink toy can be used as a stimulus to initiate the approach behavior sequence, then faded so that eventually it is no longer needed to trigger the move-forward action. This replicates the second stage in the dog training example, in which the food is moved further away from the dog on successive trials.

Two conditions — one in which the target was removed immediately after training (no fading) and one in which the target was moved gradually (fading) — were run for comparison. In the no fading condition, the pink toy was presented at a distance of 30 cm and the "move forward" command was given at the beginning of each trial. The move forward action was reinforced for 30 trials, with reinforcement being delivered after the program had moved forward 20 cm. On the thirty-first trial, only the "move forward" command was given (i.e., no toy was presented). In the fading condition, the toy was gradually faded by being presented at a distance 10 cm further on each trial for 30 trials such that by trial 30 the toy was at 330 cm. On trial 31, only the "move forward" command was given.

In the no fading condition, the program failed to move forward on the first trial during which the toy was not presented, even though the "move forward" command was given. Since the total stimulus control must be above a threshold, in this condition the command did not attain enough control during training to trigger the behavior on its own.

In the fading condition, however, the program continued to move forward during the fading procedure, as well as during the first trial during which the toy was not presented. In this condition, the gradual fading of the target allowed more associative strength to be allotted to the other stimulus: the verbal command. This, combined with a decrease in target α allowed the command to gain sufficient control such that by the thirty-first trial the command alone was a sufficient stimulus for the behavior to be emitted.

7.6. Shaping Individual Actions

In this experiment, the move-forward motor action was shaped to a particular magnitude (30 cm). First, the program was trained using the same fading procedure as in the previous experiment. This yielded an average forward distance of 20 cm. Then the target was removed, and the experimenter reinforced the program only for the largest move forward actions that it was emitting. (If no target is present, a movement distance is chosen randomly from the distribution of expected movements determined by μ and σ.) As the average magnitude of the actions increased, the experimenter continued to reinforce the program only for above average actions. When the average distance was approximately 30 cm, in order to tighten the behavior, the experimenter only reinforced actions that were close to that magnitude. By the end of 30 post-fading trials, the average distance moved was shifted from 20 cm to 30 cm.

8. Related Work

This model is related to work in a number of different areas: reinforcement learning, robot "shaping", learning through interaction with a human trainer, and learning through refinement of built-in knowledge.

8.1. Reinforcement Learning

Our model is an extension of reinforcement learning, in which an agent must find a policy mapping states to actions that maximizes some long-term measure of reinforcement. The main contribution of our work is that we address problems inherent in RL, such as dealing with large state and action spaces, by providing some built-in structure and by exploiting the domain knowledge and accurate sensory capabilities of a human trainer for behavior design and training methods. This means that our learning mechanism is less general than a standard RL technique such as Q-learning [39], but it also means that it can learn more complex tasks in a shorter amount of time. In addition, our system is capable of replicating animal learning phenomena such as fading which RL methods do not yet address.

8.2. Robot Shaping

A few researchers have implemented robot shaping systems which, although they use similar terminology, are quite different from our model. While these methods use the word "shaping" to indicate a parallel with experimental psychology, few provide any deep analysis of the animal training procedure, and in most the parallel with shaping refers to the organization of training and the fact that reinforcement, but no training examples, are provided to the learner.

Singh [36] was first in the robotics literature to use the term "shaping". In his system, complex behaviors are constructed from simple actions, that is, composite sequential decision tasks are formed by concatenating a number of elemental decision tasks. In his experiment on shaping, he trained the robot on a succession of tasks, where each succeeding task required a subset of the already learned elemental task, plus a new elemental task. After training one Q-module on an elemental task, learning was turned off for the first module and a second module was added for the composite task. Since this experiment involved the combination of elemental modules as opposed to the modulation of an existing behavior, it resembles another animal learning technique called chaining more closely than shaping. In addition, Singh does not make an effort to replicate basic aspects of animal shaping such as stimulus control. Asada [1] has also provided a modification to Q-learning that has been compared to shaping because he trains the robot on simple tasks first and moves to progressively more difficult tasks once the easier ones have been mastered.

Our work is most similar to that of Colombetti and Dorigo [11, 14] who make a point of exploiting trainer knowledge in order to speed up reinforcement learning. They also emphasize the adjustment of built-in actions through RL to tailor behaviors to specific environments. Our approach differs from theirs in several ways:

1. Their shaping procedure consists of training separate modules for different components of a behavior, and then training a separate module to coordinate them. This more closely resembles chaining, in which complex behaviors are constructed from a set of simple responses, than shaping, in which an innate behavior is modified.

2. Colombetti and Dorigo have addressed the issue of efficient search of large action spaces by employing an immediate reinforcement strategy in which every action of the robot is reinforced, either positively or negatively. This is advantageous because with feedback on every timestep, the robot usually finds the correct behavior more quickly. A drawback, however, is that immediate reinforcement is very labor intensive for the trainer. Since animal shaping procedures do not involve immediate reinforcement – the animal only receives a reward when it performs the correct behavior – our shaping methodology incorporates a delayed reinforcement strategy. Because the trainer slowly shifts the reinforcement criterion, guiding the robot through action space, the learner is able to converge on the correct behavior relatively efficiently.

3. Colombetti and Dorigo suggest that using a human trainer is not feasible because humans are inaccurate in their behavioral evaluations, and are too slow with respect to the robot's perception-action cycle. Thus, they advocate the use of a computer program as a trainer, in which case they must worry about the sensitivity of the trainer's sensors to the perception of the correct behavior on the part of the robot. While their claim about the lack of practicality in using a human trainer is probably true in the immediate reinforcement case, in the delayed

reinforcement strategy which we use this is far from the truth. One of our goals is to make the teaching of behaviors to be very straightforward, such that any lay person could easily train the robot to perform a specific task.

4. Another goal in our work is to enable these tasks to be learned while low-level reactive behaviors, such as obstacle avoidance, are intact. Because our learning program runs on top of built-in reactive behaviors [35] which incorporate no learning and which have priority, the robot is fully operational and safe from danger during new task training. Colombetti and Dorigo [11] also mention the importance of combining low-level reactive behaviors with learned behaviors. However, due to their reinforcement strategy, this can lead to problems. For example, when they were training their robot "Hamster" to "hoard food" (collect objects), they rewarded it when the distance to the "nest" decreased, and they punished it otherwise. The obstacle avoidance routine, however, interfered with this training stragegy: under their protocol the robot would be punished for avoiding obstacles if during avoidance the robot moved farther from the nest. Thus, under this system, it is necessary to train the robot while the low-level behaviors are not active, in order to keep the reinforcement programs relatively straightforward. As a result, the robot cannot be fully functional during behavior training.

5. The final difference between our work and that of Colombetti and Dorigo is the fact that, in addition to providing a novel robot learning system, we are interested in replicating phenomena associated with instrumental learning in animals, with a major goal being to understand the mechanisms underlying these processes.

8.3. Learning Through Interaction with a Human Trainer

Several researchers have introduced methods in which a human trainer provides advice to a reinforcement learning agent [17, 23, 10, 38, 22]. These systems are similar to ours in that the domain knowledge of a human trainer provides additional information to the learner, which can speed up or improve learning. However, the type of information which the human gives the learner, and the way in which the information is used, is completely different.

In all of the cited models, the trainer provides advice by suggesting actions or by giving the learner explicit rules. Thus, the agent is learning by being told. In constrast, in our model the trainer simply provides reinforcement signals when the goal behavior is performed. No explicit advice is given: the learner must determine the appropriate behavior on its own, based on the pattern of reinforcement it receives from the trainer. As we are trying to follow animal training techniques, it would not make sense to directly input high-level rules to the system, since this is something that would be well beyond the capacity of animals without language capabilities.

A second point of difference is that in all of the above systems, either the learner queries the teacher when it gets into trouble, or the teacher can input advice at any time. Our learner is set up such that the reinforcement signal is given only when the robot does the right thing for the given reinforcement schedule. In other words, the learner does not receive feedback on every timestep or with every action. This means that the human has to do less work on a given training session, making the use of the human more reasonable and less taxing.

8.4. Learning through Refinement of Built-In Knowledge

Other researchers have been concerned with establishing principles for deciding what types of behavior should be hardwired into the robot and what aspects of behavior the robot should learn from experience [24, 1, 25, 4, 5, 15].

Millán [24, 25] describes a robot learning system in which built-in basic reflexes are improved and refined through the use of RL and a neural net. These basic reflexes suggest where to search for a possible action whenever the neural network fails to generalize correctly to previous experience with the current situation. In the implementation discussed in [25], reflexes consist of move-forward actions, track object boundary, and avoid collision. These reflexes can solve the designated task (exiting a room while avoiding obstacles) without learning, but they are clumsy. After learning, the robot executes the behavior much more efficiently.

This learning system is similar to ours in that built-in behaviors are modulated with reinforcement, and that the incorporation of reflex-type behaviors allows the robot to be active from the beginning of learning. However, the modulation of behavior that can be acheived using this method is similar to only one part of our system (the shaping of individual actions). It is not clear how completely new behaviors could be shaped from innate sequences.

Other differences between our work and that of Millán are that we use delayed reinforcement from a human trainer, while he uses immediate reinforcement from a reinforcement program, and we use visual sensory input, whereas he uses sonar and IR.

Blumberg [5] describes an extensive action-selection system to which he has added some basic learning capabilities in the form of temporal difference learning [37]. This work provides a nice example of how innate knowledge is important for building complex behavior systems.

9. Discussion and Future Work

The idea of using instrumental training techniques for modifying robot behavior is appealing when one considers the complexity and diversity of behaviors that can be learned this way by animals. This richness of learning is partly due to the fact that the trainer possesses specialized skills, such as analysis of the animal's behavior based on external observation, and appropriate timing of reinforcement signals. At the same time, the actual learning process occurring within the animal's brain need not concern the trainer, thus making the training process feasible.

Our model incorporates several aspects of highly successful animal training methods based on instrumental conditioning, and replicates important animal learning phenomena such as disrimination, generalization, stimulus control, fading, and shaping of action topography. However, it is just an intial step toward a comprehensive model of conditioning. Two additional features of animal learning that we plan to add in the next version of our model, outcome devaluation and more realistic extinction, are described below.

Outcome devaluation: Much evidence exists for stimulus control being based on knowledge of the three-term relationship between the stimuli present during reinforcement (S), the action that was taken (A), and the outcome (O), rather than a simple S→R association between the stimuli and a response [13]. This can be shown by "devaluing" a reward (e.g., in rats, inducing nausea with a LiCl injection, which reduces the reward value of the type of food the rat had recently consumed) and showing that the rate of response to the controlling stimulus has diminished. If the animal were using a pure S→R strategy it would continue to emit its learned response despite the fact that the reward it was earning was no longer appetizing.

Extinction: When real animals undergo extinction (no more rewards are issued for previously-rewarded behaviors), they tend to increase both their frequency of spontaneous actions and the variability of those actions. This increase in variability can be exploited by the trainer: witholding rewards causes the animal to explore the local region of action space more widely than before, making it more likely to discover a variant that is closer to what the trainer is trying to achieve. A similar idea for varying the distribution of generated actions is described in [24].

In addition to providing a model of animal learning, our system seeks to diminish some of the scaling problems that exist in more general reinforcement learning systems. This is done through the incorporation of structure into the learner that allows it to be guided through state and action space, as opposed to blind exploration of every possible state and action combination. Guidance is accomplished in three ways:

Modification of innate behavior sequences. We do not assume that everything must be learned from scratch; instead we rely on built-in structure. Much evidence exists for the notion that animals have hardwired behavior sequences, and that there are "natural units" of behavior [19, 20]. Conventional in ethological research on learning [2], and later accepted by neobehaviorist experimental psychologists [6], is the idea that conditioned behavior is synthesized out of instinctive acts. Adult animals come to a learning situation with a structured response hierarchy.[2] The goal of shaping is to use this response hierarchy to the trainer's advantage by molding pre-existing responses into more desired behaviors.

One stimulus controls behavior. By the end of training the robot on a specific task, a small number of stimuli have gained complete control over the behavior, such that no other stimuli are important. Thus, whenever the robot perceives the controlling stimulus, it will execute the behavior no matter what the other aspects of the state are. This is modulated, of course, by the low-level reactive behaviors that are constantly running in order to prevent the robot from getting into danger. The idea of stimulus control puts more power into the hands of the trainer, in that the trainer must decide upon appropriate stimuli and must command the robot to do tasks. At the same time, stimulus control can reduce the combinatorial problem which results when a mapping

[2]Whether this response hierarchy is truly innate (i.e., a result of genetics) or learned earlier in life is not relevant. We simply assume that the animal, or robot in this case, has some pre-existing response structure; it doesn't matter where it came from.

between every possible aspect of a state and the appropriate action must be produced.

Utilization of a human trainer. The main principle behind shaping, in the animal learning literature and in our model, is that the reinforcement contingencies are changed slowly from training trial to training trial. Instead of choosing actions randomly and then updating their likelihood of occurrence based on environmental consequences, the robot learns by being reinforced by the trainer for close approximations to the desired behavior, with gradually tightening tolerances. Essentially, the robot is *guided through action space*, being reinforced for closer and closer approximations to, and eventually converging on, the desired behavior.

In sum, we have presented a novel robot learning system that is based on principles from instrumental conditioning in animals. These principles are incorporated into a computational model of shaping which is based on behavior editing to transform pre-existing actions into novel behaviors. In this paper, we show how the model allows a human trainer to teach a mobile robot several tasks, all derived from the same underlying behavior sequence. Aspects of our approach, such as use of a human trainer and training through reinforcement, have driven work in several different research areas. We are attempting to extend these efforts by building a learning system that replicates as many established instrumental learning phenomena as possible. It is our hope that a model which incorporates these various aspects of animal learning will yield a richer and more robust method for training mobile robots, and at the same time will provide a new perspective on the mechanisms underlying animal learning.

Acknowledgments

The research was funded by National Science Foundation grant IRI-9530975. We thank Joseph O'Sullivan for technical assistance with the Amelia robot and Reid Simmons for making Amelia available to us.

References

[1] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Purposive behavior acquisition for a real robot by vision-based reinforcement learning. *Machine Learning*, 23(2-3):279–303, 1996.

[2] S.A. Barnett. *Modern Ethology.* Oxford University Press, 1981.

[3] D. A. Baxter, D. V. Buonomano, J. L. Raymond, D. G. Cook, F. M. Kuenzi, T. J. Carew, and J. H. Byrne. Empirically derived adaptive elements and networks simulate associative learning. In *Neural Network Models of Conditioning and Action*, pages 13–52. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[4] B. Blumberg. Action-selection in hamsterdam: Lessons from ethology. In *Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior*, Brighton, 1994.

[5] B. M. Blumberg, P. M. Todd, and Pattie Maes. No bad dogs: Ethological lessons for learning in hamsterdam. In *Proceedings of the 4th International Conference on the Simulation of Adaptive Behavior*, 1996.

[6] K. Breland and M. Breland. The misbehavior of organisms. *American Psychologist*, 16:681–684, 1961.

[7] P.L Brown and H.M. Jenkins. Auto-shaping of the pigeon's keypeck. *Journal of the Experimental Analysis of Behavior*, 11:1–8, 1968.

[8] T. J. Bussey, J. L. Muir, and T. W. Robbins. A novel automated touchscreen procedure for assessing learning in the rat using computer graphic stimuli. *Neuroscience Research Communications*, 15(2):103–109, 1994.

[9] CCI. *The CCI Program.* Canine Companions for Independence, Santa Rosa, CA, 1995. Informational page available at http://grunt.berkeley.edu/cci/cci.html.

[10] J. A. Clouse and P. E. Utgoff. A teaching method for reinforcement learning. In *Proceedings of the Ninth Conference on Machine Learning.* Morgan Kaufmann, 1992.

[11] M. Colombetti, M. Dorigo, and G. Borghi. Behavior analysis and training: A methodology for behavior engineering. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 26(3):365–380, 1996.

[12] A. Dickinson. Instrumental conditioning. In N. J. Mackintosh, editor, *Handbook of Perception and Cognition. Volume 9*. Academic Press, Orlando, FL, 1995.

[13] Anthony Dickinson. Actions and habits: the development of behavioral autonomy. *Philosophical Transactions of the Royal Society of London, Series B*, 308:67–78, 1985.

[14] M. Dorigo and M. Columbetti. Robot shaping: Developing autonomous agents through learning. *Artificial Intelligence*, 70(2):321–370, 1994.

[15] G. L. Drescher. *Made-Up Minds*. The MIT Press, Cambridge, MA, 1991.

[16] C. R. Gallistel. *The Organization of Action*. Lawrence Erlbaum Associates, Hillsdale,NJ, 1980.

[17] D. Gordon and D. Subramanian. A multistrategy learning scheme for agent knowledge acquisition. *Informatica*, 17:331–346, 1994.

[18] R. E. Hampson, C. J. Heyser, and S. A. Deadwyler. Hippocampal cell firing correlates of delayed-match-to-sample performance in the rat. *Behavioral Neuroscience*, 107(5):715–739, 1993.

[19] E. Hearst and H.M. Jenkins. *Sign tracking: The stimulus-reinforcer relation and directed action*. The Psychonomic Society, Austin, 1975.

[20] H.M. Jenkins and B.R. Moore. The form of the autoshaped response with food or water reinforcers. *Journal of the Experimental Analysis of Behavior*, 20:163–181, 1973.

[21] L. Pack Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[22] L.-J. Lin. Self-improving reactive agents based on reinforcement learning, planning, and teaching. *Machine Learning*, 8:293–321, 1992.

[23] R. Maclin and J. W. Shavlik. Creating advice-taking reinforcement learners. *Machine Learning*, 22(1-2-3):251–281, 1996.

[24] J. del R. Millán. Learning efficient reactive behavioral sequences from basic reflexes in a goal-directed autonomous robot. In *From Animals to Animats 3: Proceedings of the Third*

International Conference on Simulation of Adaptive Behavior, pages 266–274, Cambridge, MA, 1994. MIT Press.

[25] J. del R. Millán. Rapid, safe, and incremental learning of navigation strategies. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 26(3):408–420, 1996.

[26] J.M. Pearce and G. Hall. A model for Pavlovian learning: Variations in effectiveness of conditioned but not unconditioned stimuli. *Paychological Review*, 87(6):532–552, 1980.

[27] S.M. Pellis, D.P. O'Brien, V.C. Pellis, P. Teitelbaum, D.L. Wolgin, and S. Kennedy. Escalation of feline predation along a gradient from avoidance through play to killing. *Behavioral Neuroscience*, 102(5):760–777, 1988.

[28] Simon Perkins and Gillian Hayes. Robot shaping – principles, methods, and architectures. In *Workshop on Learning in Robots and Animals, AISB '96*, 1996.

[29] K. Pryor. *Lads Before the Wind.* Harper and Row, New York, 1975.

[30] J. L. Raymond, D. A. Baxter, D. V. Buonomano, and J. H. Byrne. A learning rule based on empirically derived activity-dependent neuromodulation supports operant conditning in a small network. *Neural Networks*, 5(5):789–803, 1992.

[31] R. A. Rescorla and A. R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy, editors, *Classical Conditioning II: Theory and Research.* Appleton-Century-Crofts, New York, 1972.

[32] G. S. Reynolds. *A Primer of Operant Conditioning.* Scott, Foresman, 1968.

[33] F. Rosenblatt. *Principles of Neurodynamics.* Spartan, New York, 1962.

[34] R. Simmons. Structured control for autonomous robots. *IEEE Transactions on Robotics and Automation*, 10(1):34–43, 1994.

[35] Reid Simmons, Richard Goodwin, Karen Haigh, Sven Koenig, and Joseph O'Sullivan. A modular architecture for office delivery robots. In *The First International Conference on Autonomous Agents*, Feb 1997.

[36] S.P. Singh. Transfer of learning across sequential tasks. *Machine Learning*, 8:323–339, 1992.

[37] R. S. Sutton and A. G. Barto. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170, 1981.

[38] P. Utgoff and J. Clouse. Two kinds of training information for evaluation function learning. In *Proceeding of the Ninth National Conference on Artificial Intelligence (AAAI-91).* AAAI Press, 1991.

[39] C. J. C. H. Watkins. *Learning from Delayed Rewards.* PhD thesis, Cambridge University, Cambridge, England, 1989.

Figure Captions

Figure 1: A basic object pursuit behavior sequence.

Figure 2: Interruption of the object pursuit sequence by food reinforcement from the trainer.

Figure 3: After a number of interrupted trials the importance of the sight of food for triggering the object pursuit sequence decreases. At the same time, since the "beg" command has been consistently presented, the associative strength between the commanded and the rewarded behavior increases.

Figure 4: The sight of food as a controlling stimulus has been faded, and now the dog will beg in response to the command alone.

Figure 5: Basic behavior sequence for the robot, defined in terms of states and links. Transition links connect states at the same level of abstraction; sub-state activation links connect an abstract state with one or more concrete realizations.

Figure 6: An example precondition.

Figure 7: Status of pre- and postconditions for the approach target state when the robot is 100 cm from the target. Aggregate control of the three postconditions is less than 0.1. (Below threshold of 0.5.)

Figure 8: Status of pre- and postconditions for the approach-object state after the robot has moved within reach of the target. Aggregate control of the three postconditions is 0.6. (Above threshold of 0.5.)

Figure 9: Amelia being trained to put toys in the bin.

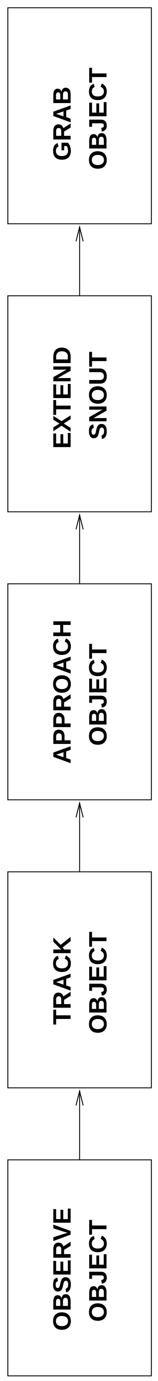

Figure 1: A basic object pursuit behavior sequence.

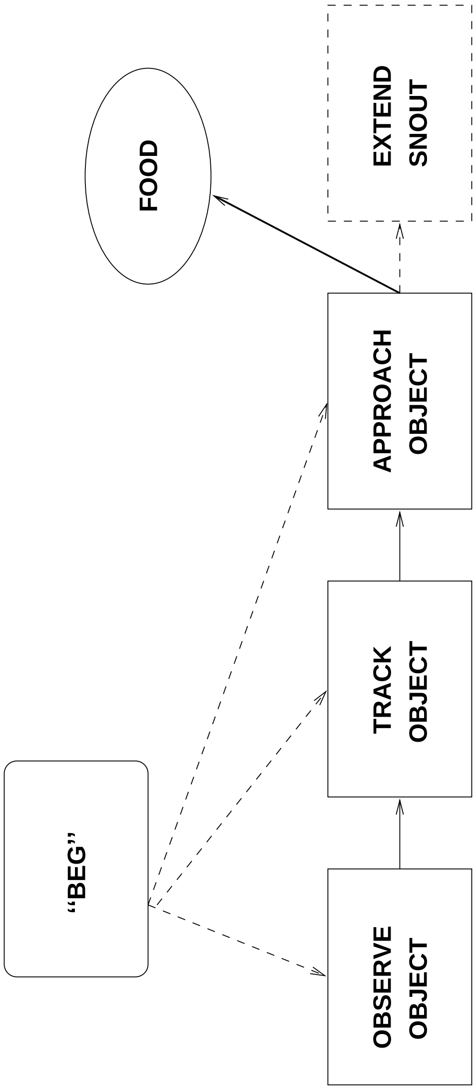

Figure 2: Interruption of the object pursuit sequence by food reinforcement from the trainer.

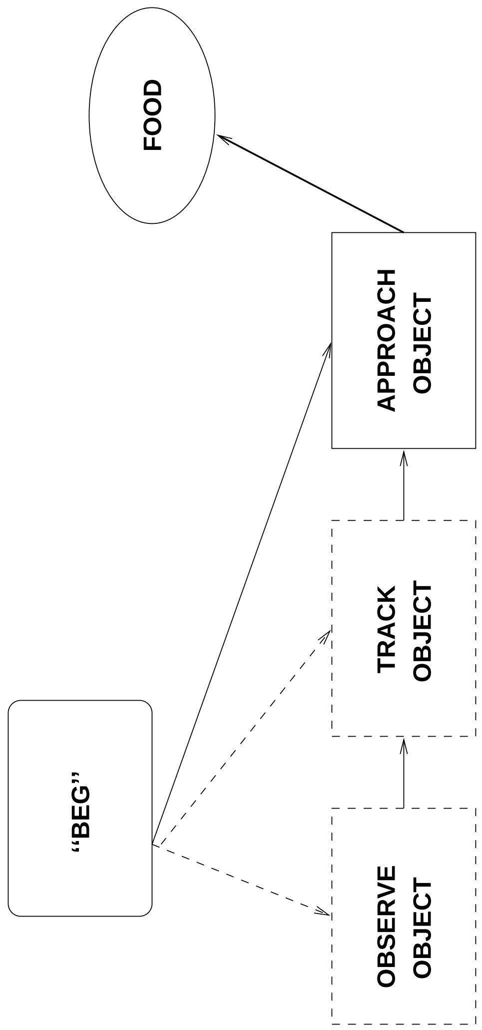

Figure 3: After a number of interrupted trials the importance of the sight of food for triggering the object pursuit sequence decreases. At the same time, since the “beg” command has been consistently presented, the associative strength between the command and the rewarded behavior increases.

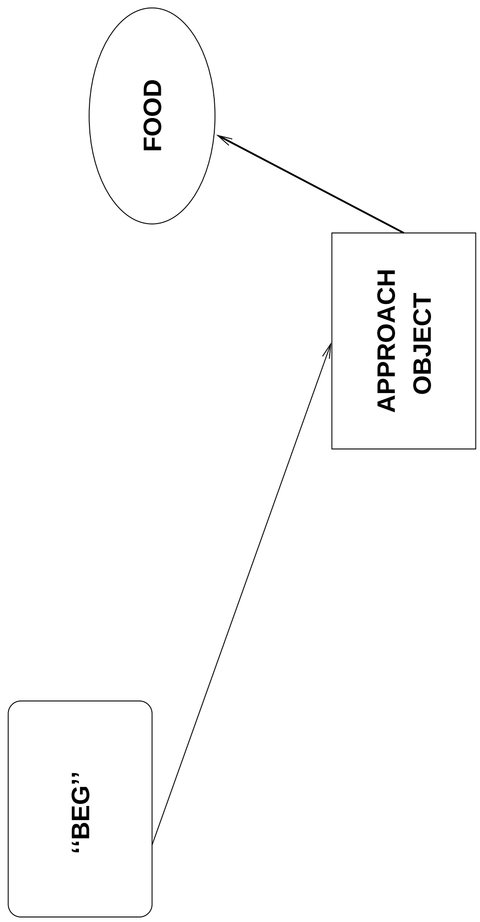

Figure 4: The sight of food as a controlling stimulus has been faded, and now the dog will beg in response to the command alone.

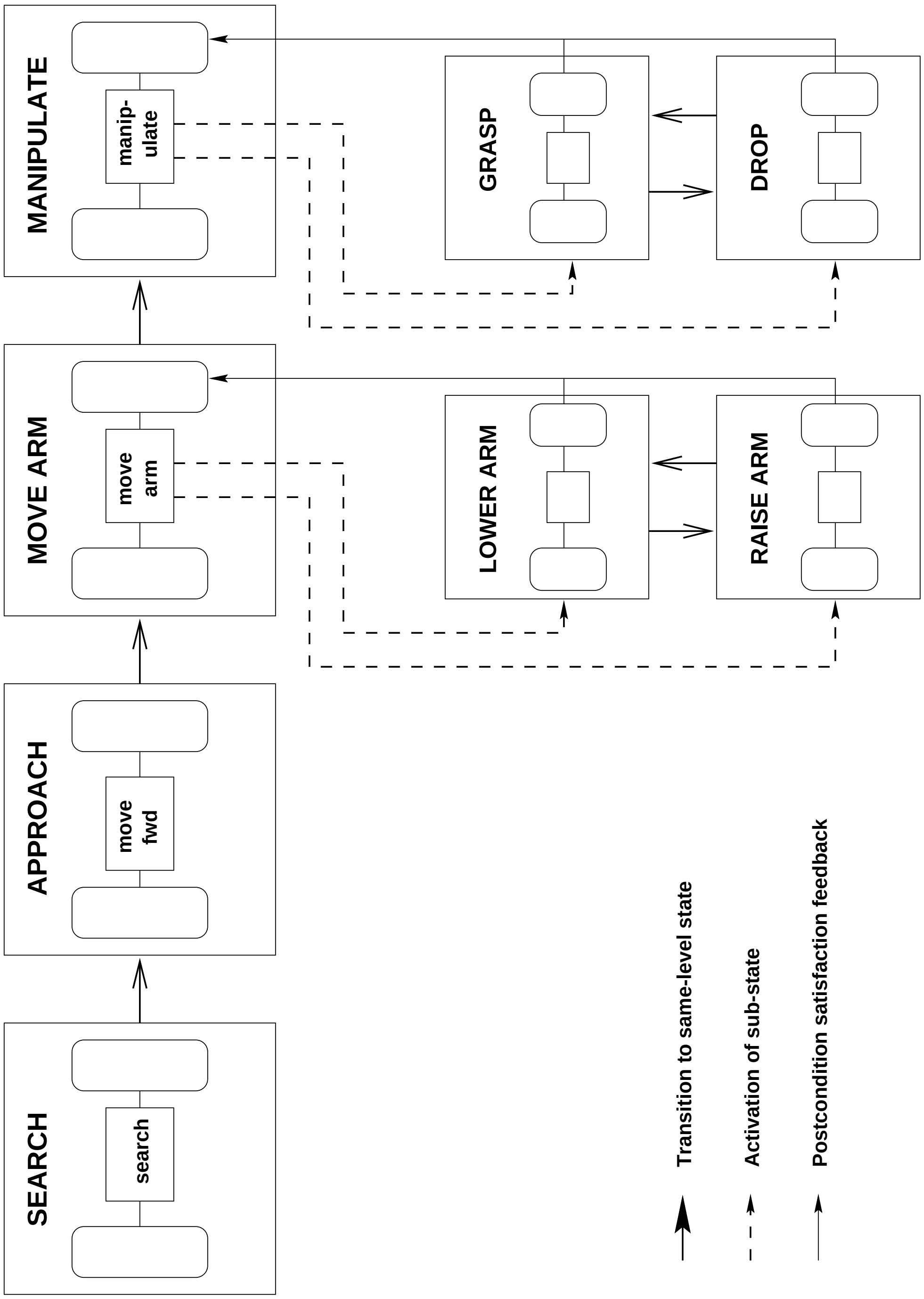

Figure 5: Basic behavior sequence for the robot, defined in terms of states and links. Transition links connect states at the same level of abstraction; sub-state activation links connect an abstract state with one or more concrete realizations.

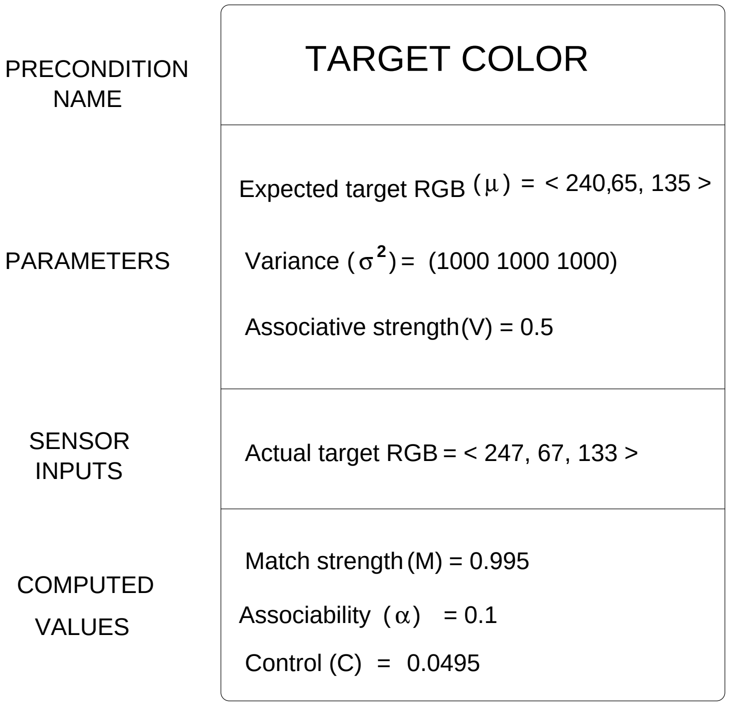

Figure 6: An example precondition.

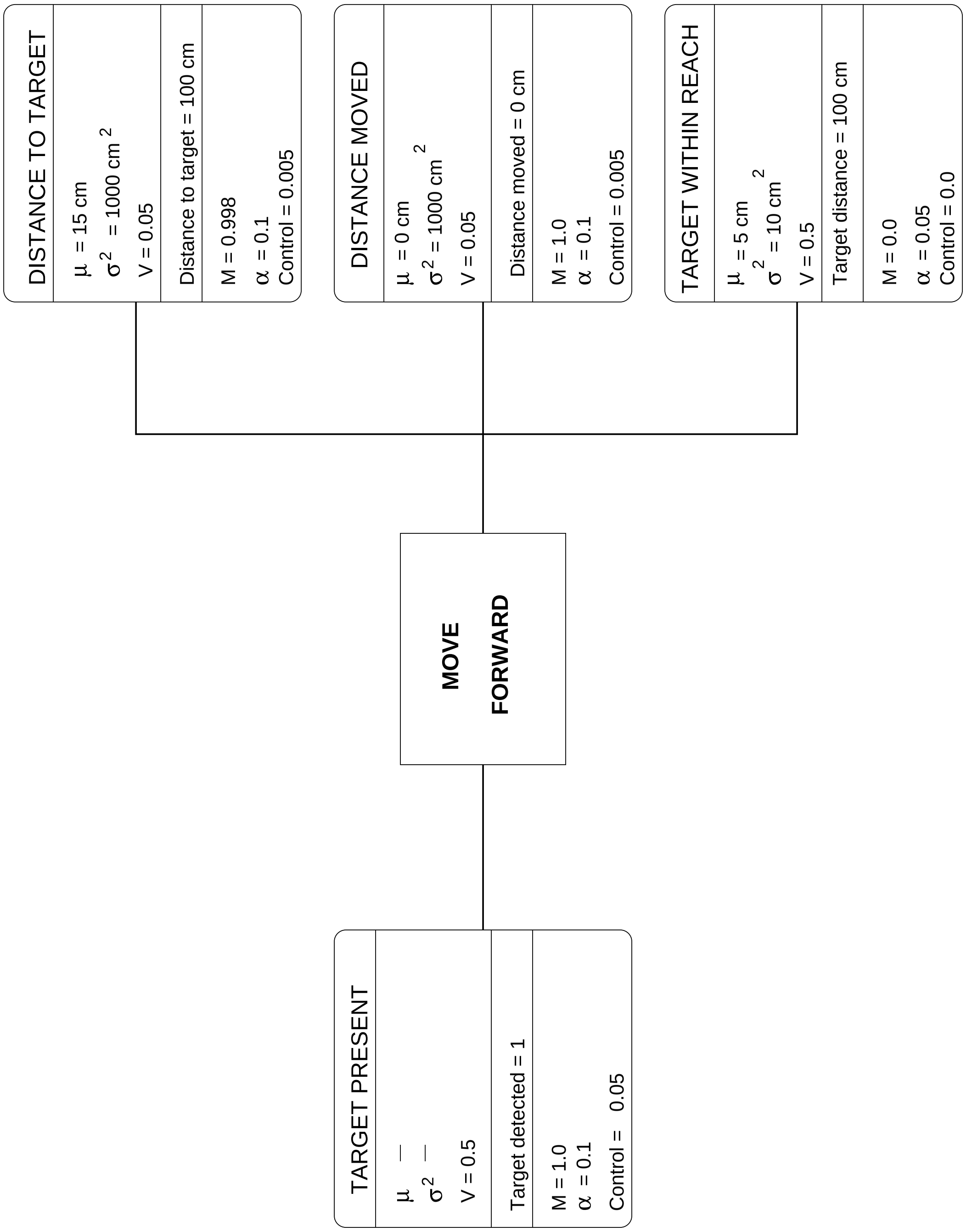

Figure 7: Status of pre- and postconditions for the approach target state when the robot is 100 cm from the target. Aggregate control of the three postconditions is less than 0.1. (Below threshold of 0.5.)

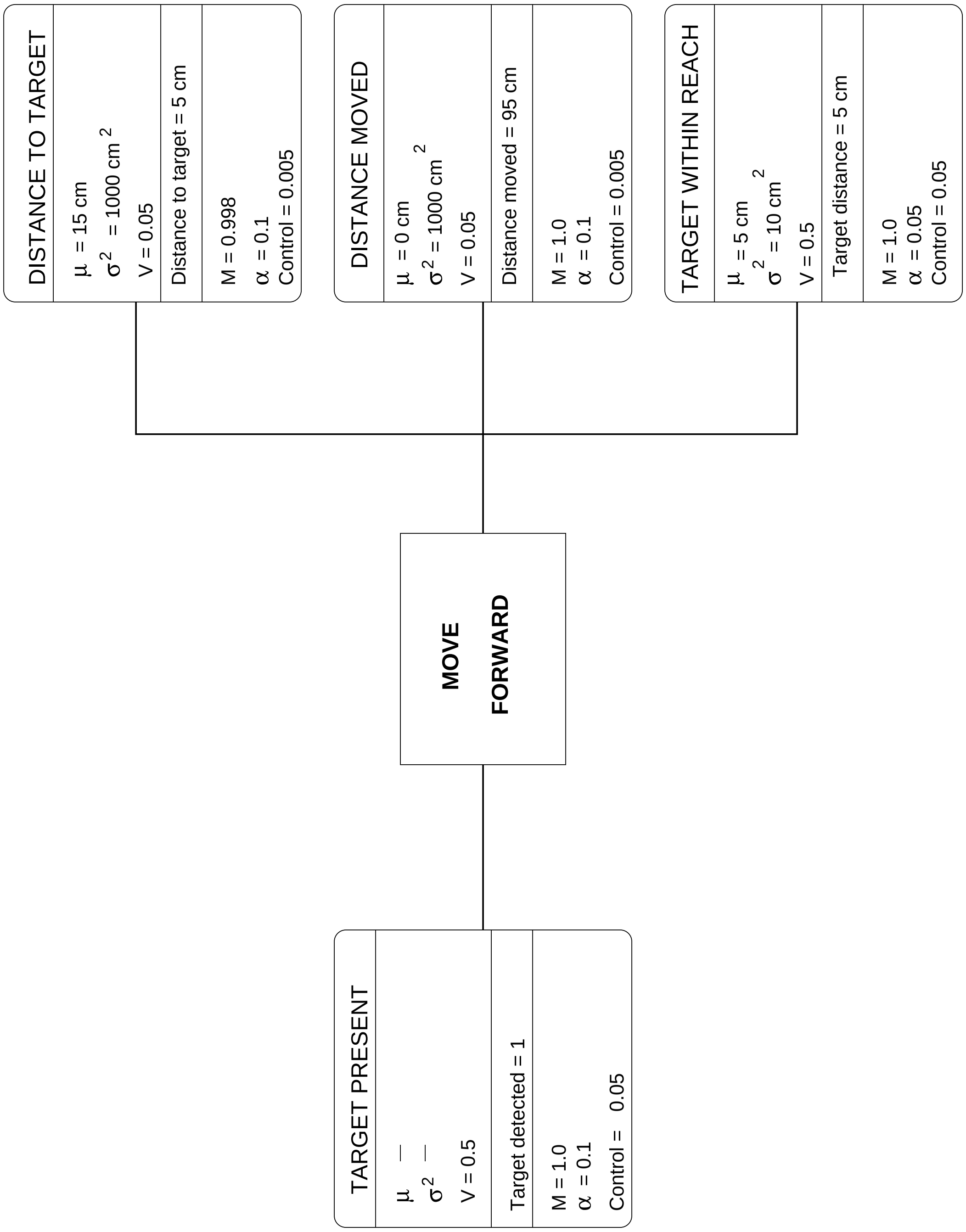

Figure 8: Status of pre- and postconditions for the approach-object state after the robot has moved within reach of the target. Aggregate control of the three postconditions is 0.6.(Above threshold of 0.5.)

Figure 9: Amelia being trained to put toys in the bin.