Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors

Yael Niv [1], Daphna Joel [1], Isaac Meilijson [2] and Eytan Ruppin [3]

[1] Department of Psychology
Tel-Aviv University, Tel-Aviv 69978, Israel
yaeln@cns.tau.ac.il, *djoel@post.tau.ac.il*

[2] School of Mathematical Sciences
Tel-Aviv University, Tel-Aviv 69978, Israel
isaco@math.tau.ac.il

[3] School of Computer Sciences & Sackler School of Medicine
Tel-Aviv University, Tel-Aviv 69978, Israel
ruppin@math.tau.ac.il

Corresponding author:
Yael Niv
Tel: +972-3-6407864, +972-3-6490264
Fax: +972-3-6409113
Email: yaeln@cns.tau.ac.il

Abstract

Reinforcement learning is a fundamental process by which organisms learn to achieve goals from their interactions with the environment. Using Evolutionary Computation techniques we evolve (near) optimal neuronal learning rules in a simple neural network model of reinforcement learning in bumblebees foraging for nectar. The resulting neural networks exhibit efficient reinforcement learning, allowing the bees to respond rapidly to changes in reward contingencies. The evolved synaptic plasticity dynamics give rise to varying exploration/exploitation levels and to the well-documented choice strategies of risk aversion and probability matching. Risk-averse behavior is evolved even in a risk-less environment, and in contrast to existing theories in economics and game theory, it is shown to be a direct consequence of optimal reinforcement learning, without requiring additional assumptions such as the existence of a non-linear subjective utility function. Our results are corroborated by a rigorous mathematical analysis, and their robustness in real-world situations is supported by experiments in a mobile robot. Thus we provide a biologically founded, parsimonious and novel explanation for risk aversion and probability matching.

Keywords: Reinforcement Learning, Risk Aversion, Evolutionary Computation, Dopamine, Neuromodulation, Heterosynaptic Plasticity

Running head: Evolution of Reinforcement Learning

1 Introduction

Reinforcement learning (RL) is a process by which organisms learn from their interactions with the environment to achieve a goal (Sutton & Barto, 1998). In RL, learning is contingent upon a scalar reinforcement signal which provides evaluative information about how good an action is in a certain situation, without providing an instructive supervising cue as to which would be the preferred behavior in the situation. Behavioral research indicates that RL is a fundamental means by which experience changes behavior in both vertebrates and invertebrates, as most natural learning processes are conducted in the absence of an explicit supervisory stimulus (Donahoe & Packard-Dorsel, 1997). Several brain regions have been implicated in RL, including the midbrain dopaminergic neurons of the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) in rats and primates, and their target areas in the basal ganglia (e.g. Graybiel & Kimura, 1995; Houk & Wise, 1995; Schultz, 1998). A computational understanding of neuronal reinforcement learning will enhance the understanding of learning processes in the brain, and can contribute widely to the design of autonomous artificial learning agents.

RL has attracted ample attention in computational neuroscience, yet a fundamental question regarding the underlying mechanism has not been sufficiently addressed, namely, **what are the optimal synaptic learning rules for maximizing reward in RL?** In this paper, we use Evolutionary Computation techniques to derive the **optimal neuronal learning rules that give rise to efficient reinforcement learning in uncertain environments.** We then investigate the behavioral strategies, which emerge as a result of optimal RL.

RL has been demonstrated and studied extensively in foraging bees, thus we have chosen bee foraging as a model system for studying synaptic learning rules for RL. Real (1991, 1996) showed that when foraging for nectar in a field of blue and yellow artificial flowers, bumblebees exhibit efficient RL, rapidly switching their preference for flower type when reward contingencies were switched between the flowers. The bees also manifested risk-averse behavior: in a situation in which blue flowers contained $2\mu l$ sucrose solution, and yellow flowers contained $6\mu l$ sucrose in $\frac{1}{3}$ of the flowers, and zero in the rest, $\sim 85\%$ of the bees' visits were to the blue constant-rewarding flowers, although the mean return from both flower types was identical. Risk-sensitive and risk-averse choice behavior has also been demonstrated extensively in other animals (see Kacelnik & Bateson, 1996, for a review).

Bees foraging for nectar are faced with highly variable ecological conditions during the course of the year and in different habitats: parameters such as the weather, the season, and competition, all affect the availability of rewards from different kinds of flowers. In such an environment rapid learning is crucial for successful foraging, as the foraging individual cannot be prepared genetically for the ecological conditions of a particular habitat (Menzel & Muller, 1996). An uncertain environment also implies an "armed-bandit" type scenario, in which the bee collects food and information simultaneously (Greggers & Menzel, 1993). The foraging bee's choices are guided not only by the search for food, but also by the search for information regarding the content of different food sources. This implies a tradeoff between exploitation and exploration (Wilson, 1996), as the bee's choices directly affect the "training examples" which it will encounter through the learning process. Thus in an "armed bandit" situation a bee must devise a policy for choosing between exploiting available knowledge or exploring for more information, at every trial.

In a previous neural network (NN) model, Montague, Dayan, Person, and Sejnowski (1995) simulated bee foraging in a 3D arena of blue and yellow flowers, based on a neurocontroller modelled after an identified interneuron (VUMmx1) in the honeybee suboesophogeal ganglion (Hammer, 1993). This neuron's activity represents the reward value of gustatory stimuli, and similar to midbrain dopaminergic neurons of primates, it is activated by unpredicted rewards and by reward predicting stimuli, and is not activated by predicted rewards (Hammer, 1997). In their model, this neuron is modeled as a linear unit P, which receives visual information regarding changes in the percentages of yellow, blue and neutral colors in the visual field, and computes a prediction error. According to P's output the bee decides whether to continue flight in the same direction, or to change heading direction randomly. Upon landing, a reward value is applied to the network, according to the subjective utility of the nectar content of the chosen flower (Harder & Real, 1987), and the synaptic weights of the networks are updated according to a special anti-Hebbian-like learning rule in which the postsynaptic factor selects the direction of change (Montague, 1997). As a result, the values of the weights come to represent the expected rewards from each flower type. The behavior of the decision unit P in this model is similar to the firing patterns characteristic of dopaminergic neurons in the SNc and VTA of primates as have been recorded in a monkey performing a delayed-match-to-sample task (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997), and is thus consistent with the hypothesis that these dopaminergic neurons compute a prediction error (Schultz, 2000).

While this model replicates Real's foraging results and provides a basic and simple NN architecture to solve RL tasks, many aspects of the model, first and foremost the handcrafted synaptic learning rule, are arbitrarily specified and their optimality with respect to RL questionable. Towards this end, we use a generalized and parameterized version of this model in order to evolve optimal synaptic learning rules for RL (with respect to maximizing nectar intake) using a genetic algorithm (Mitchell, 1997). Evolutionary computation techniques are especially suitable for RL problems, as they involve an artificial decision-making agent acting in an environment in order to achieve a goal. The solution is "evolved" in much the same way as the biological solutions have been evolved by natural selection - through a "parallel search" for effective solutions in a large population of individuals.

In contrast to common evolutionary computation applications that involve NNs with evolvable synaptic weights or architectures (Ackley & Littman, 1991; Floreano & Mondada, 1996; Nolfi, Elman, & Parisi, 1994), we set upon the task of evolving the network's neuronal learning rules. Previous attempts at evolving neuronal learning rules have used heavily constrained network dynamics and very limited sets of learning rules (Baxter, 1992; Chalmers, 1990; Floreano & Mondada, 1998; Fontanari & Meir, 1991), or evolved only a subset of the learning rule parameters (Unemi et al., 1994). We define a general framework for evolving Hebbian learning rules, which essentially encompasses all heterosynaptic and monosynaptic Hebbian learning rules and also allows for complex neuromodulatory interactions of gating of synaptic plasticity. Via the genetic algorithm we select bees based solely on their nectar-gathering ability in a changing environment. The uncertainty of the environment ensures that efficient foraging can only be a result of learning throughout lifetime, thus efficient learning rules are evolved.

To avoid the possible confusion of terms, we make a distinction between the notions of heterosynaptic learning (Dittman & Regehr, 1997; Schacher, Wu, & Sun, 1997; Vogt & Nicoll, 1999) and neuromodulation of plasticity (Bailey, Giustetto, Huang, Hawkins, & Kandel, 2000; Fellous & Linster, 1998). The classic monosynaptic Hebbian learning rule is an activity-dependent learning rule in which a synapse is updated only when there is activity both in the presynaptic and the postsynaptic neurons. In contrast, heterosynaptic Hebbian learning allows for activity-independent modification of synapses such that a synapse can also be updated when only the presynaptic or postsynaptic component has been active, and more generally, even when neither have been active. We term this rule "heterosynaptic" modification as it allows for the firing of a neuron to effect all its synapses, regardless of the activity of the other neurons connected to them. Neuromodulation of synaptic plasticity further enhances the learning rule by allowing a three-factor interaction in the learning process: through neuromodulation the activity of a neuron can gate the plasticity of a synapse between two other neurons. Providing the neuromodulatory neuron has fired, the synapse can be updated (according to the heterosynaptic learning rule pertaining to the pre- and postsynaptic neurons). Otherwise, plasticity in the synapse is shut off. Both heterosynaptic plasticity and neuromodulatory gating of synaptic plasticity have been demonstrated in neural tissues (Bailey et al., 2000; Dittman & Regehr, 1997; Fellous & Linster, 1998; Schacher et al., 1997; Vogt & Nicoll, 1999), and have been recognized to increase the computational complexity of synaptic learning (Bailey et al., 2000; Fellous & Linster, 1998; Wickens &

Kötter, 1995). By allowing for heterosynaptic learning and neuromodulation of plasticity, we define a very large search space in which the genetic algorithm can search for optimal synaptic learning rules.

In the following section we describe the model and the evolutionary dynamics. Section 3 reports the results of our simulations: In 3.1 we describe the successful evolution of RL. Section 3.2 describes the evolved synaptic update rules, and their influence on the exploration/exploitation tradeoff of the foraging bees. In section 3.3 we analyze the foraging behavior resulting from the learning dynamics, and find that when tested in a new environment, our evolved bees manifest risk-averse behavior. Although this choice strategy was not selected for, we rigorously prove that risk aversion emerges directly from optimal RL, in contrast to the conventional explanations of risk-averse behavior which rely on the existence of subjective utility functions. The strength of the evolved learning mechanism is further demonstrated in section 3.4, which describes the emergence of probability matching behavior. This behavior, which was previously thought to result from competition for food resources, is shown to emerge in a non competitive scenario as a result of the learning rule dynamics alone. Section 3.5 describes a minirobot implementation of the model, aimed at assessing its robustness. We conclude with a discussion of our results in section 4.

2 The Model

A simulated bee-agent flies in a 3D arena, over a 60x60 flower patch composed of randomly scattered yellow and blue squares representing two types of flowers. A bee's life consists of 100 trials. Each trial begins with the bee placed in a random location above the flower patch and with a random heading direction. The bee starts its descent from a height of $\sim$ 10 units above the flower patch, and advances in steps of 1 unit that can be taken in any downward direction (360° horizontal, 90° vertical). The bee views the world through a cyclopean eye (10° cone view), and in each timestep decides whether to maintain the current heading direction or to reorient randomly, based on the visual input. Upon landing (the field has no boundaries, and the bee can land on a flower or outside the flower patch on neutral-colored ground), the bee consumes any available nectar in one timestep, and another trial begins. **The evolutionary goal (the fitness criterion) is to maximize nectar intake.**

FIGURE 1 should come about here

In the neural network controlling the bee's flight (Figure 4a), which is an extension of Montague et al's (1995) network, three modules ("regular", "differential" and "reward") contribute their input via synaptic weights, to a linear neuron P. The regular input module reports the percentage of the bee's field of view filled with yellow $[X_y(t)]$, blue $[X_b(t)]$ and neutral $[X_n(t)]$. The differential input module reports temporal differences of these percentages $[X_i(t) - X_i(t-1)]$. The reward module reports the actual amount of nectar received from a flower $[R(t)]$ in the nectar-consuming timestep (in this timestep it is also assumed that there is no new input $[X_i(t) = 0]$), and zero during flight. Note that, in contrast to Montague et al. (1995) **we do not incorporate any form of nonlinear utility function** with respect to the reward. Thus P's continuous-valued output is:

$$P(t) = R(t) + \sum_{i \in \text{regular}} W_i X_i(t) + \sum_{j \in \text{differential}} W_j [X_j(t) - X_j(t-1)] \tag{1}$$

The bee's action is determined according to the output $P(t)$ using Montague et al's (1995) probabilistic action function (Figure 4b):

$$p(\text{change direction}) = \frac{1}{1 + \exp[m \cdot P(t) + b]} \tag{2}$$

where m and b are real-valued evolvable parameters.

During the bee's "lifetime" the synaptic weights of the regular and differential modules are modified via a heterosynaptic Hebb learning rule of the form:

$$\Delta W_i(t) = \begin{cases} \eta[A * V_i(t)P(t) + B * V_i(t) + C * P(t) + D] & \text{dependencies met} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where η is a global learning rate parameter, $V_i(t)$ and $P(t)$ are the pre-synaptic and the post-synaptic values respectively, W_i their connection weight, and $A - D$ are real-valued evolvable parameters. In addition, learning in one module can be dependent on another module (dashed arrows in Figure 4a), such that if module M depends on module N, M's synaptic weights will be updated according to equation (3) only if module N's neurons have fired, and if it is not dependent, the weights will be updated on every timestep. In this respect, a dependency on the reward module is satisfied when the reward neuron fires, and dependencies on the regular and differential modules are satisfied neuron-wise, i.e. when a neuron fires, the synapses connected to the respective (same color) neurons in the dependent module can be updated. Synapses of a module dependent on two other modules can only be updated when satisfying both dependency conditions. Thus the bee's "brain" is capable of a non-trivial neuromodulatory gating of synaptic plasticity.

The simulated bee's genome (Figure 4c) consists of a string of 28 genes, each representing a parameter governing the network architecture or learning dynamics. Fifteen genes specify the bee's brain at time of "birth" (before the first trial): 7 boolean genes determine whether each synapse exists or not; 6 real-valued genes (range [-1,1]) specify the initial weights of the regular and differential module synapses (the synaptic weight of the reward module is clamped to 1, effectively scaling the other network weights); and two real-valued genes specify the action-function parameters m (range [5,45]) and b (range [0,5]). Thirteen remaining genes specify the learning dynamics of the network: The regular and differential modules each have a different learning rule specified by 4 real-valued genes (parameters $A - D$ of equation (3), range [-1,1]); The global learning rate of the network η is specified by a real valued gene; and four boolean genes specify dependencies of the visual input modules on each of the other two modules.

The optimal gene values were determined using a genetic algorithm. A first generation of bees was produced by randomly generating 100 genome strings. Each bee performed 100 trials independently (no competition) and received a fitness score according to the average amount of nectar gathered per trial. To form the next generation, fifty pairs of parents were chosen (with returns) with a bee's fitness specifying the probability of it being chosen as a parent. Each two parents gave birth to two offsprings, which inherited their parents' genome after performing recombination and adding random mutations. Mutations were performed by adding a uniformly distributed value in the range of [-0.1,0.1] to 2% of the real-valued genes, and reversing 0.2% of the boolean genes. Recombination was performed via a uniform crossover of the genes ($p = 0.25$, genewise). It is important to note that there was no Lamarkian inheritance - learned weights were not passed on to offsprings. One hundred offsprings were thus created, and once again tested in the flower field. This process continued for a large number of generations.

3 Results

3.1 Evolution of Reinforcement Learning

To promote the evolution of learning, bees were evolved in an "uncertain" world: In each generation one of the two flower types was randomly assigned as a constant-yielding high-mean flower (containing $0.7\mu l$ nectar), and the other a variable-yielding low-mean flower ($1\mu l$ nectar in $\frac{1}{5}$th of the flowers and zero otherwise). The reward contingencies were switched between the two flower types in a randomly chosen trial during the second or third quarter of each bee's life. Evolutionary runs under this condition typically show one of two types of fitness curves: successful runs, defined as runs in which reward-dependent choice behavior is successfully evolved, are characterized by two distinct evolutionary jumps (Figure 4a).

Unsuccessful runs which produce behavior that is not dependent on rewards, show only the first jump (Figure 4b). This jump is due to the acquisition of a negative neutral synapse responsible for a "do not land on neutral-colored ground" rule.

FIGURE 2 should come about here

About half of the evolutionary runs were successful. Figure 4a shows the mean value of several of the bees' genes in the last generation of each of five successful runs. The second evolutionary jump characteristic of successful runs is due to the almost simultaneous evolution of 8 genes governing the network structure and learning dependencies, which are essential for producing efficient learning in the bees' uncertain environment: All successful networks have a specific architecture which includes only four synapses (the regular neutral, differential blue and differential yellow and the reward synapse), as well as a dependency of the differential module on the reward module, conditioning modification of these synapses on the presence of reward. Agents which have almost all the crucial alleles, but are missing one or two, are nevertheless unsuccessful (Figure 4b). Thus we find that in our framework, only a network architecture similar to that used by Montague et al. (1995) can produce above-random foraging behavior, supporting their choice as an optimal one. However, **our optimized networks utilize a heterosynaptic learning rule different from the monosynaptic rule used by Montague et al., giving rise to several important behavioral strategies.**

FIGURE 3 should come about here

3.2 Exploration/Exploitation Tradeoff and Heterosynaptic Modification

In order to understand the evolved learning rule, we examined the foraging behavior of individual bees from the last generation of successful runs. In general, the bees manifest efficient reinforcement learning, showing a marked preference for the high-mean rewarding flower, with a rapid transition of preferences after the reward contingencies are switched between the flower types. The values of the synaptic weights are also indicative of learning based on reward contingencies, as they follow the rewards expected from each flower.

FIGURE 4 should come about here

A more detailed inspection of the behavior of bees from different evolutionary runs reveals that the bees differ in their degree of exploitation of the high-rewarding flowers versus exploration of the other flowers (Figure 4). These individual differences in the foraging strategies employed by the bees, result from an interesting relationship between the micro-level Hebb rule coefficients and the exploration/exploitation tradeoff characteristic of the macro-level behavior. According to the dependencies evolved, learning (synaptic updating) occurs only upon landing, and we can analyze the evolved heterosynaptic learning rule of the differential module as follows: In the common case, upon landing the bee sees only one color, thus all inputs are zero except the differential input corresponding to the color of the chosen flower which is, in the absence of new visual input in the landing step: $X_i(t)-X_i(t-1) = 0-1 = -1$. Thus the output of P in this step is:

$$P(t) = R(t) + (-1) \cdot W_{\text{chosen}}(t) = R(t) - W_{\text{chosen}}(t) \tag{4}$$

Therefore, the synaptic update rule for the differential synapse corresponding to the chosen flower color is:

$$\Delta W_{\text{chosen}}(t+1) = \eta[(A-C) \cdot (-1) \cdot (R(t) - W_{\text{chosen}}(t)) + (D-B)] \tag{5}$$

leading to an effective monosynaptic coefficient of $(A - C)$, and a general weight decay coefficient $(D - B)$. For the other differential synapses the synaptic update rule is:

$$\Delta W_j(t+1) = \eta[C \cdot (R(t) - W_{\text{chosen}}(t)) + D] \tag{6}$$

TABLE 1 should come about here

Table 1 summarizes the values of the coefficients of the heterosynaptic learning rule of the differential module, as they evolved in five different successful evolutions. The synaptic weight of the chosen color is affected by the values of $(A-C)$ and $(D-B)$. The large negative values of $(A-C)$ found specify an anti-Hebbian learning rule (i.e. a anti-correlation learning rule) for the chosen flower. This is a result of the negative presynaptic value whilst landing. The values of D and B further modulate the synaptic strength: in trials in which no prediction error is encountered (i.e. the postsynaptic value is near zero in the learning step), a negative value of $(D-B)$ results in weakening of the synapse while a positive value results in strengthening of the synapse. The synaptic weights corresponding to colors which were not chosen are only influenced heterosynaptically, by the values of C and D. A positive value of C results in a global strengthening of these synapses when the postsynaptic value is positive and a global weakening when the postsynaptic value is negative, i.e. a generalization of the "good/bad surprise" which was encountered, to the other colors. A negative value of C will produce the opposite effect. A positive or negative value of D results in a global strengthening or decay of all the synapses at every learning step.

Thus the heterosynaptic learning rules evolved cover a large range of synaptic dynamics which affect all the synapses of the differential module at every learning step, and influence the bee's exploration/exploitation tradeoff. A positive value of D results in "spontaneous" strengthening of competing synapses, leading to an exploration-inclined bee. A positive C value further enhances this behavior. Conversely, negative values of C and D will result in a declining tendency to visit competing flower types, leading to exploitation-inclined behavior.

3.3 Emergence of Risk Aversion

A prominent strategy exhibited by the evolved bees is risk aversion. Figure 4a shows the choice behavior of previously evolved bees, tested in a new environment in which the mean rewards of the two kinds of flowers are identical. Although the situation does not call for any flower preference, as on average the bee would receive the same nectar amount regardless of its choice strategy, the bees consistently prefer the constant-rewarding flower. This result is not a consequence of evolution in conditions in which the variably-rewarding flower was also the low-mean rewarding flower, as we have found that risk-averse behavior can also be evolved without introducing any risk to the environment during the evolutionary process: Bees evolved in an environment containing two constant-rewarding flowers yielding different amounts of nectar, also exhibit risk-averse behavior when tested in a variable-rewarding flower scenario. Figure 4b shows the choice preferences of a new population of bees evolved in an environment which contained two constantly-rewarding flower types, one yielding $0.8\mu l$ nectar, and the other $0.3\mu l$. The environment was an uncertain one, in that the flower color which yielded the higher reward was randomly chosen in each generation, and the reward contingencies were switched between the two flower types in a random timestep within the second or third quarter of the bees' lifetime, similar to the previoulsly described evolutionary runs. After completing 500 generations of evolution, 40 bees from the last generation were tested in the risk aversion scenario in which both flowers yielded the same mean reward, but one was variably-rewarding. Confronted with variably rewarding flowers for the first time, the bees showed a pronounced preference for the constant rewarding flower type (Figure 4b). **Thus we can observe emergent risk-averse behavior in bees evolved in a risk-less environment.**

FIGURE 5 should come about here

Risk aversion has been studied extensively in the fields of economics and game theory. The risk-averse behavior which has been observed in many choice scenarios in animals as well as in humans, has traditionally been accounted for by hypothesizing the existence of a non-linear concave "utility function" for reward. In bees, for instance, such a subjective utility function for nectar can result from a concave relationship between nectar volume and net energy intake, between net energy intake and fitness, or between the actual and perceived nectar volume (Harder & Real, 1987; Smallwood, 1996). In contradistinction to the conventional explanations of risk aversion common in the fields of economics and game theory, our model does not include a non-linear utility function. **What hence brings about risk-averse behavior in our model?** Corroborating previous numerical results (March, 1996), we prove analytically that this foraging strategy is a direct consequence of Hebbian learning dynamics in an armed-bandit-like RL situation.

During a trial, a bee makes a series of choices regarding its flight direction, in order to choose which flower to land on. As the bee does not learn (i.e. there is no synaptic plasticity) during flight, all the choices throughout one trial are influenced by the same weight values. Thus under a certain rewarding regime, i.e. in between changes in reward contingencies, the bee's stochastic foraging decisions can be modelled as choices between a variable-rewarding (v) and a constant-rewarding (c) flower, based on synaptic weights W_v and W_c.

For simplicity, we examine the case of simple monosynaptic anti-Hebbian learning: In this case, the synaptic update rule is the well-known temporal difference (TD) learning rule (Sutton & Barto, 1998)

$$\Delta W(t) = \eta(R(t) - W(t-1)). \tag{7}$$

The synaptic weights are in effect a "memory" mechanism, as they are a function of the rewards previously obtained from each of the two flower types. W_v, representing the reward expected from the variable flower, is thus an exponentially weighted average of $[v_1, v_2, \cdots]$, the previous rewards obtained from (v):

$$W_v = W_v(\eta) = \eta(v_t + (1-\eta)v_{t-1} + (1-\eta)^2 v_{t-2} + \cdots) \tag{8}$$

W_c, as an exponentially weighted average of rewards obtained from the constant-rewarding flower type, is constant.

Bearing this notion in mind, the Appendix provides a mathematical proof of the emergence of risk aversion. We compute f_v, the frequency of visits to variably rewarding flowers, and show that it is a function of $p_v(W_v)$, defined as the probability of choosing (v) in a trial in which the synaptic weight corresponding to the variably rewarding flower is W_v. We prove that W_v, as a function of the learning rate, is risk-ordered, such that for higher learning rates $W_v(\eta)$ is riskier than for lower learning rates. We then use the mathematical definition of riskiness to show that under relatively mild assumptions regarding the bee's choice function, the frequency of visits to variably rewarding flowers is lower for higher learning rates, than for lower learning rates, and is ordered by learning rate. Finally we show that the risk order property of $W_v(\eta)$ always implies risk-averse behavior, i.e. **for every learning rate, the frequency of visits to the variable flower (f_v) is less than 50%, further decreasing under higher learning rates.** Our simulations corroborate these analytical results (Figure 4c).

Intuitive insight to these results can be gained by observing that in the learning process, the bee makes its decisions based on finite time-windows, and does not compute the long-term mean reward obtained from each flower. This is even more pronounced with high learning rates such as those evolved (~ 0.8). With such a learning rate, after landing on an empty flower of the variable-rewarding type, the bee updates the reward expectation from this flower type (i.e. updates the corresponding synaptic weight according to the evolved heterosynaptic Hebb update rule) to near zero, and as a result, prefers the constantly rewarding flower, from which it constantly expects (and receives) a reward of $\frac{1}{2}\mu l$. As long as the bee chooses the constant-rewarding flower, it will not update the expectation from the variable-rewarding flower, which will remain near zero. Even after an occasional "exploration" trial in which a visit to the variable flower yields a high reward, the preference for this flower will be short lived, lasting only until the next unrewarded visit. Note that rapid learning such has been evolved here, is essential for obtaining high fitness in a highly variable environment (Menzel & Muller, 1996), and such abnormally high learning rates have been hypothesized by Real (1991), and were also used in Montague et al.'s (1995) model. Nevertheless, the above mathematical analysis shows that even with low learning rates, as long as the bee is a reinforcement-learning bee (its learning rate greater than zero), it will manifest risk-averse behavior.

3.4 Emergence of Probability Matching Behavior

Another notable strategy by which bumblebees (and other animals) optimize choice in armed-bandit situations is probability matching. When faced with flowers offering similar rewards but with different probabilities, bees have been shown to choose the different flower types by a ratio that matches the reward ratio of the flowers (Greggers & Menzel, 1993; Kaesar, Rashkovich, Cohen, & Shmida, Submitted). This seemingly "irrational" behavior with respect to optimization reward intake was explained as an Evolutionary Stable Strategy (ESS) for the individual forager, when faced with competitors (Thuijsman, Peleg, Amitai, & Shmida, 1995). In a multi-animal competitive setting, probability matching produces an Ideal Free Distribution (IFD) in which the average intake of food is the same at all food sources, and no animal can improve its payoff by feeding at another source. Using evolutionary computation techniques, Seth (1999) evolved battery-driven agents which competed for two different battery refill sources, and showed that indeed matching behavior emerges in a multi-agent scenario, while when evolved in isolation, agents choose only the high probability refill source.

FIGURE 6 should come about here

Surprisingly, our evolved bees also demonstrate probability matching behavior. Figure 4 shows the previously evolved bees' performance when tested in matching experiments in which all flowers yield $1\mu l$ nectar, but with different reward probabilities. In both conditions, the bees show near-matching behavior, preferring the high-probability flower to the low-probability one, by a ratio that closely matches the reward probability ratios. This emergent behavior is not a result of competitive conditions (as the bees were evolved and tested in isolation), and is also a direct result of the evolved reinorcement learning dynamics: Due to the high learning rate, the fluctuating weights representing the expected yield from each flower will essentially move back and forth from zero to one. When both are zero, the two flowers are chosen randomly, but the high yielding flower has a greater chance of yielding reward, after which its weight will be updated to 1, and this flower is preferred to the other. When both weights are 1, the less-yielding flower has a greater chance of having its weight updated to zero, again resulting in preference for the high-yielding flower. Thus, in contradistinction to previous accounts, probability matching behavior can be evolved in a non-competitive setting, again as a direct consequence of optimal RL.

3.5 Robot Implementation

In order to assess the robustness of the evolved RL algorithm, we implemented it in a mobile mini-robot by letting the robot's actions be governed by a NN controller similar to that evolved in successful bees, and by having its synaptic learning dynamics follow the previously evolved RL rules. A Khepera mini-robot foraged in a 70X35cm arena whose walls were lined with flowers, viewing the arena via a low-resolution CCD camera (200x200 pixels), moving at a constant velocity and performing turns according to the action function (eq. 2) in order to choose flowers, in a manner completely analogous to that of the simulated bees. The NN controller was identical to that evolved for the simulated bees, except that it received no "neutral" inputs. All calculations were performed in real-time on a Pentium-III 800Mhz computer (256Mb RAM) in tether mode. Moving with continuous speed and performing all calculations in real-time, the foraging robot exhibited rapid reinforcement learning and risk-averse behavior, analogous to that of the simulated bees (Figure 4). Thus the algorithms and behaviors evolved in the virtual bees' simulated environment using discrete time-steps hold also in the different and noisy environment of real foraging mini-robots operating in continuous time.

FIGURE 7 should come about here

4 Discussion

The interplay between learning and evolution has been previously investigated in the field of evolutionary computation. Much of this research has been directed towards elucidating the relationship between evolving traits (such as synaptic weights) versus learning them (Ackley & Littman, 1991; Hinton & Nowlan, 1987). A relatively small amount of research has been devoted to the evolution of the learning process itself, most of which was constrained to choosing the appropriate learning rule from a limited set of predefined rules (Baxter, 1992; Chalmers, 1990; Floreano & Mondada, 1996). In this work we show for the first time that optimal learning rules for RL in a general class of armed bandit situations, can be evolved in a general Hebbian learning framework. The evolved learning rules are by no means trivial, as they are heterosynaptic and employ synaptic plasticity modulation.

As a result of the evolved learning rules, several complex foraging behaviors emerge, demonstrating the strength of evolutionary computation as a neuroscience research methodology that links together phenomena on the neuronal and behavioral levels. **We show that the fundamental macro-level strategies of risk aversion and probability matching are a direct result of the micro level synaptic learning dynamics**, which also control the tradeoff between exploration and exploitation. These behavioral strategies have not been explicitly evolved, but emerge in the model as "side-effects" of optimal learning, making additional assumptions conventionally used to explain these behaviors redundant. **Furthermore, in this work we not only show that risk-averse behavior can exist in the absence of a non-linear utility function, but also show that this behavior can emerge spontaneously in a risk-free environment.** Risk aversion is a direct consequence of adaptation to a changing, uncertain world, that induces a form of learning in which the estimation of expected rewards is biased, as it takes the more recent sampling into account more strongly. When acting in an environment in which intra-flower risk is added to the inter-flower uncertainty, this bias is expressed as risk-averse behavior. This result is important not only to the fields of evolutionary computation and animal learning theories, but also to the fields of economics and game theory, as it provides a new perspective on the well-studied paradigm of risk-sensitive behavior.

The prevailing accounts of risk aversion in the fields of economics and game theory are based on the notion of non-linear utility functions, which were introduced into rational choice theory models in order to explain the apparent divergence from rationality embedded in risk-sensitive behavior. Utility functions were later assumed to result from biological/energetic considerations, an explanation that usually involves postulating and quantifying a large number of parameters (for instance see Harder and Real (1987) for the energetic derivation of the utility function for nectar postulated for bumble-bees). It should be stressed that, although widely accepted as an explanation for risk aversion, there is no direct evidence for the existence of subjective utility functions other than the behavioral manifestation of risk-sensitive behavior itself. The existence of the assumed underlying subjective utility function, and the shape of the function, are assessed not by an independnt measure of subjective utility, but according to the resulting risk-averse behavior. In contrast, in our model the well-documented phenomenon of risk aversion characteristic of choice behavior in many species, emerges directly from the learning dynamics of a simple neural network model based on an architecture identified in brains of bumblebees and mammals alike. We have shown empirically and theoretically that a finite memory, which is a direct consequence of Hebbian learning, is sufficient to account for risk aversion in any model of temporal difference learning which involves a learning rate greater than zero. As non-linear utility functions are abstract concepts that have not been proven to exist biologically, omitting them from the model has the effect of producing a more generally applicable model which includes less species-specific parameters and functions, and can therefore account for risk-averse behavior in many reinforcement learning situations and in different species.

According to our mathematical analysis and simulation results (see Figure 4c), there is a direct relationship between the model bee's learning rate and the resulting risk-averse behavior. Based on this we can predict that individual differences between bees in the observed risk-averse behavior, will be correlated with the learning rate of the bees. This prediction can be empirically tested by assesing the learning rate of individual bees based on a learning task, and then measuring the amount of risk aversion displayed by the bees in a foraging task similar to that described in this work.

Keeping in mind that the model we have described is an abstract learning model and therefore should be related to biological learning mechanisms with caution, the evolved learning architecture can be used to shed light on the biological implementation of reinforcement learning. Thus, the significance of using heterosynaptic learning rules in our model should be noted: Through the learning rules evolved, a synapse can be modified even when its corresponding flower type was not chosen. This allows for non-trivial interactions between flower types when forming predictions as to the reward expected from each flower. For example, the expectation from one flower type can be updated as a function of the disappointment or surprise encountered when choosing the other flower type. Heterosynaptic learning rules have been used in computational modelling of synaptic plasticity, and recognized as contributing to the complexity of the learning process, but surprisingly few studies have directly explored this phenomenon in the brain. Evidence from cerebellar (Dittman & Regehr, 1997) and hippocampal (Vogt & Nicoll, 1999) synapses shows that heterosynaptic plasticity indeed occurs in the brain, and can affect the spiking patterns of neurons through interactions between adjacent synapses. In-vitro recordings in Aplysia have directly demonstrated heterosynaptic facilitation of synapses from two presynaptic neurons onto a common postsynaptic target (Schacher et al., 1997). We show here that heterosynaptic plasticity can exert a pronounced effect on the behavior of the organism, and specifically illustrate this effect on the tradeoff between exploration and exploitation characteristic of armed bandit situations.

The learning mechanism described in this work can be closely related to the "adaptive critic" described by Sutton and Barto in the actor-critic architecture frequently used to implement RL in artificial agents (Barto, 1995; Sutton, 1988). Furthermore, although embedded in a very simple artificial neural-network, the learning rules we have studied are biologically plausible and can be related to learning dependent on the dopaminergic system in the basal ganglia (Houk, Adams, & Barto, 1995). In the actor-critic model, an actor sub-network learns to perform actions so as to maximize the weighted sum of future rewards, which is computed at every timestep by a critic sub-network (Barto, 1995). The critic is adaptive, in that it learns to predict the weighted sum of future rewards based on the current sensory input and the actor's policy, by means of an iterative process in which it compares its own predictions to the actual rewards obtained by the acting agent. The learning rule used by the adaptive critic is the temporal differnce (TD) learning rule (Sutton, 1988) in which the error between two adjacent predictions (the TD error) is used to update the critic's weights. In our model the synaptic weights come to represent the expected rewards, and as the inputs are differential, the activity of P represents an ongoing comparison between the expected reward in subsequent timesteps. As in the critic model (Barto, 1995), this comparison provides the error measure by which the network updates its weights and learns to better predict future rewards.

Why, then, is learning in this model restricted by the intermodule dependencies only to the landing step, while learning in the classic adaptive critic model occurs at every step? The answer to this lies in the simplicity of the model and the learning task. The classic adaptive-critic uses the error measure in order to learn to better predict the consequences of every input scenario, based on similarity of the current sensory input to previously encountered (and learned) inputs (Barto, 1995). This learning scheme is capable of dealing with complex situations in which outcomes depend on a long sequence of actions, but the final outcome of actions can not be easily predicted by the sensory inputs in every timestep. Our model bee deals with a much simpler problem, as in every timestep it can directly assess the final outcome of flying in its current direction. The inputs available to our model bee are not the primary sensory input from the retina, but a pre-processed input consisting of the percentages of reward-predicting features (flower colors) in the primary input. As a result of the construct of the learning task, these inputs are not only sufficient to drive the learning process to generate valid predictions, but once a prediction is generated (according to the reward obtained), this prediction can be completely generalized to all the different input cases encountered by the bee during flight. As given a set of weights (which define the prediction of the network), a consistent prediction can be made for any combination of color percentages in the visual field, there is no prediction error to learn from during the bee's flight. The prediction itself can be updated only when the bee lands and encounters (or does not encounter) an actual reward, which can be compared to the predicted reward.

Several studies have suggested that dopaminergic neurons in the basal ganglia may constitute a biological implementation of a TD reinforcement signal (Montague, Dayan, Nowlan, Pouget, & Sejnowski, 1993; Montague et al., 1996; Schultz et al., 1997), and together with the striatum may implement an actor-critic learning model (Houk et al., 1995; Schultz, 1998). The output of unit P in our model, as in that of Montague et al's (1995) model, quite accurately captures the essence of the activity patterns of midbrain dopaminergic neurons (Montague et al., 1996; Schultz et al., 1997) in primates and rodents, and the corresponding octopaminergic neurons in bees (Hammer, 1997; Menzel & Muller, 1996). These neurons, originating in the SNc and VTA of primates and rodents, presumably project a widespread prediction-error signal to many cortical and subcortical areas, including the striatum (Schultz, 1998; Schultz et al., 1995; Waelti, Dickinson, & Schultz, 2001). In corticostriatal synapses, this signal is implicated in mediating learning by reinforcement, by neuromodulating synaptic plasticity (providing a "now learn" signal which allows synaptic plasticity and long-term memory formation in areas which were active at the same time or immediately before a rewarding input was encountered (Bailey et al., 2000; Wickens & Kötter, 1995)).

The similarity between DA-dependent plasticity in corticostriatal synapses (Kimura & Matsumoto, 1997; Suzuki, Miura, Mishimura, & Aosaki, 2001; Wickens & Kötter, 1995) and the dependencies evolved in our model should be noted, However, this comparison should be done with caution, as ours is a very simplified model. We have found that efficient RL is dependent upon a three factor Hebbian learning rule, in which the synaptic weights are updated as a function of a neuromodulating reward signal, as well as the presynaptic and postsynaptic factors. Our demonstration of the optimality of this learning rule to RL has bearing on the computational function of DA-dependent plasticity in the BG. Since the current model reflects mainly the critic module of the actor-critic framework and consists only of a very simplistic actor (the probabilistic action function), future work which will focus on elaborating the actor component of the model, will undoubtly increase the relevance of the model to learning in the basal ganglia.

In summary, the significance of this work is three-fold: on the one hand we show the strength of simple evolutionary computation models in evolving fundamental processes such as reinforcement learning, and on the other hand we show that optimal reinforcement learning can directly explain complex behaviors such as risk aversion and probability matching, without need for further assumptions. This is done by means of a very simple (and thus easily analyzed) model, which nonetheless has important bearing on learning in biological neuronal circuits.

APPENDIX

Mathematical Analysis - Risk Aversion is Ordered by learning rate

We consider the bee's long-term choice dynamics as a sequence of N cycles, each choice of (v) beginning a cycle. Let $n_i \geq 0$ be the number of visits to constant flowers in the i'th cycle. The frequency f_v of visits to (v) is determined (via Birkhoff's Ergodic theorem (Breiman, 1968), an extension to dependent variables of the Strong Law of Large Numbers) by the expected number of visits to (c) in a typical cycle $[E(n)]$:

$$f_v = \lim_{N\to\infty} \frac{N}{N + \sum_{i=1}^{N} n_i} = \lim_{N\to\infty} \frac{1}{1 + \frac{1}{N}\sum_{i=1}^{N} n_i} = \frac{1}{1 + E(n)} \tag{9}$$

As W_c is constant, the bee's choices are only a function of W_v, and we can define the bee's choice function as $p_v(W_v)$, the probability of choosing (v) in a trial in which the synaptic weight corresponding to the variably rewarding flower is W_v. Thus given W_v, $[n_i + 1]$ is geometrically distributed with $p_v(W_v)$, giving:

$$E(n) = E[E(n|W_v)] = E\left[\frac{1}{p_v(W_v)} - 1\right] = E\left[\frac{1}{p_v(W_v)}\right] - 1 \tag{10}$$

and so

$$f_v = \frac{1}{E[\frac{1}{p_v(W_v)}]} \tag{11}$$

According to the mathematical definition of riskiness which comes from theories of second degree stochastic dominance (Hardy, Littlewood, & Polya, 1934), for X and Y with a finite equal mean, Rothschild and Stiglitz (1970) show the following three statements to be equivalent:

(i) $EU(X) \geq EU(Y)$ for every concave function U for which these expectations exist

(ii) $E[\max(X - x, 0)] \leq E[\max(Y - x, 0)]$ for all $x \in \Re$

(iii) There exists on some probability space two random variables X and Z such that $Y = X + Z$ and $E(Z|X) = 0$ with probability 1.

Statement (i) provides the mathematical definition of riskiness, i.e. X is less risky than Y if (i) is true, as for every concave utility function the mean subjective reward obtained from X is greater than that obtained from Y so every risk averter would prefer X to Y. Statement (ii) is an easier condition to check when determining which of two random variables is riskier. Statement (iii) is a mathematically equivalent definition of riskiness which we will use later in our analysis to prove that the bee is always risk-averse. According to this:

Lemma: If $X, X_1, X_2, X_3, \ldots$ are identically distributed (not necessarily independent) random variables with a finite mean, $Y = \sum_{i=1}^{\infty} \alpha_i X_i$ (where $\vec{\alpha}_i$ is a probability vector) is less risky than X.

Proof:

$$\sum \alpha_i X_i - x = \sum \alpha_i (X_i - x) \leq \sum \alpha_i [\max(X_i - x), 0] \tag{12}$$

Since the right-hand side is non-negative,

$$\max[\sum \alpha_i X_i - x, 0] \leq \sum \alpha_i [\max(X_i - x), 0] \tag{13}$$

Now taking expectations of both sides

$$\begin{aligned} E[\max(\sum \alpha_i X_i - x), 0] &\leq \sum \alpha_i E[\max(X_i - x), 0] = \\ &= \sum \alpha_i E[\max(X - x), 0] = \\ &= E[\max(X - x), 0] \qquad \therefore \end{aligned} \tag{14}$$

As a corollary, we shall prove that exponential smoothers such as $W_v(\eta)$ are risk-ordered such that a lower learning rate β leads to less risk aversion than a higher learning rate α $(0 < \beta < \alpha < 1)$.

Lemma: Let $W_v(\eta)$ be an exponentially weighted average of identically distributed variables V_i $(i = 1, 2, 3, \ldots)$ as in equation (8), then $W_v(\alpha)$ is riskier than $W_v(\beta)$ for every $0 < \beta < \alpha < 1$

Proof: Let us define $W_v^{(k)}(\alpha)$ identically distributed (not independent) variables as the following:

$$\begin{aligned}
W_v^{(1)}(\alpha) &= \alpha v_1 + \alpha(1-\alpha)v_2 + \alpha(1-\alpha)^2 v_3 + \cdots \\
W_v^{(2)}(\alpha) &= \alpha v_2 + \alpha(1-\alpha)v_3 + \alpha(1-\alpha)^2 v_4 + \cdots \\
&\vdots \\
W_v^{(k)}(\alpha) &= \alpha v_k + \alpha(1-\alpha)v_{k+1} + \alpha(1-\alpha)^2 v_{k+2} + \cdots \qquad (15)
\end{aligned}$$

Let us now choose a specific probability vector $(\vec{\alpha}_i)$ as following:

$$\alpha_1 = \frac{\beta}{\alpha}\ ;\ \alpha_n = \alpha_1(\alpha-\beta)(1-\beta)^{n-2}\ (n \geq 2) \qquad (16)$$

We then have:

$$\begin{aligned}
\sum_{k=1}^{\infty} \alpha_k W_v^{(k)}(\alpha) &= \frac{\beta}{\alpha} \cdot \alpha \cdot [v_1 + (1-\alpha)v_2 + (1-\alpha)^2 v_3 + \cdots \\
&+ (\alpha-\beta)v_2 + (\alpha-\beta)(1-\alpha)v_3 + (\alpha-\beta)(1-\alpha)^2 v_4 + \cdots \\
&+ (\alpha-\beta)(1-\beta)v_3 + (\alpha-\beta)(1-\beta)(1-\alpha)v_4 + \\
&+ (\alpha-\beta)(1-\beta)(1-\alpha)^2 v_5 + \cdots] = \\
&= \beta[v_1 + (1-\beta)v_2 + (1-\beta)^2 v_3 + \cdots] = W_v(\beta) \qquad (17)
\end{aligned}$$

Thus $W_v(\beta)$, as a weighted sum of $W_v^{(k)}(\alpha)$, is less risky than $W_v(\alpha)$. $\therefore$

Now tying this to equation (11) and to statement (i) of the Rothschild and Stiglitz (1970) theorem, if $\phi(\cdot) = \frac{1}{p_v(\cdot)}$ is convex (and so $-\frac{1}{p_v(\cdot)}$ is concave), then

$$E\left(-\frac{1}{p_v(W_v(\beta))}\right) \geq E\left(-\frac{1}{p_v(W_v(\alpha))}\right) \tag{18}$$

$$\Rightarrow f_v(\alpha) = \frac{1}{E\left(\frac{1}{p_v(W_v(\alpha))}\right)} \leq \frac{1}{E\left(\frac{1}{p_v(W_v(\beta))}\right)} = f_v(\beta) \tag{19}$$

And the bee will display **ordered** risk-averse behavior: The higher the learning rate, the lower is the frequency of visits f_v to the (v) flowers. Convexity of $\frac{1}{p_v(\cdot)}$ is a mild assumption as for every concave increasing p_v, $\frac{1}{p_v}$ is strictly convex, so convexity will also be preserved under minor departures from concavity of p_v.

According to statement (iii) of the Rothschild and Stiglitz (1970) theorem, if Y is obtained from X by further fair gambling, then X is less risky than Y. Thus when both flower types yield the same mean reward, W_v is riskier than W_c. From this follows that even with low learning rates, since p_v is symmetric with respect to W_v and W_c (i.e. $p_v(W_c) = \frac{1}{2}$), when both flower types reward with the same mean, the frequency f_v is always less than $\frac{1}{p_v(W_c)} = \frac{1}{2}$, and **the bee is always risk-averse**.

Acknowledgements

We thank the Telluride Workshop on Neuromorphic Engineering (2000) for providing the setting and equipment for the mini-robot experiments. We are grateful to Tamar Keasar for the valuable contribution of ideas and experimental data from probability matching experiments in bees.

Figure Captions:

Figure 1

(a) The bee's neural network controller. The weights $W_i(t)$ of the regular and differential modules are modifiable. **(b) The bee's action function.** Probability of reorienting direction of flight as a function of $P(t)$ for different values of parameters m, b. **(c) The "genome" sequence of the simulated bee.**

Figure 2

Typical fitness scores of a successful run (a) and an unsuccessful run (b) of 500 generations. Solid line - mean fitness, dotted line - maximum fitness in each generation.

Figure 3

Mean value of several genes in the last generation of (a) successful and (b) unsuccessful runs. Each sub-figure shows the mean value of one gene in the last generation of each of five runs. Genes shown (from left to right): Top row - the learning rate gene and the two action function parameters m and b. Middle row - the boolean genes governing the existence of the different synapses: regular blue, regular yellow and regular neutral input synapses, differential blue, differential yellow and differential neutral input synapses, and the reward input synapse. Bottom row - boolean genes determining the dependencies of the regular module on the differential module and on the reward module, and the dependencies of the differential module on the regular module and the reward module.

Figure 4

Preference for blue flowers for two different bees from the last generation of different successful runs, averaged over 40 test bouts, each consisting of 100 trials. Blue is the initial constant-rewarding high-mean flower. Reward contingencies were switched between flower types at trial 50. Hebb rule coefficients for the "exploiting" bee (c) are $A = -0.82, B =$

$0.15, C = 0.24, D = -0.04$ and for the "exploring" bee (d) are $A = -0.92, B = 0.39, C = 0.16, D = 0.25$.

Figure 5

(a) Risk aversion - Preference for blue flowers in 100 test trials averaged over 40 previously evolved bees, now tested in conditions different from those they were evolved in. Although both flower types yield the same mean reward (blue - $\frac{1}{2}\mu l$ nectar, yellow - $1\mu l$ in half the flowers, contingencies switched at trial 50), there is a marked preference for the constant-yielding flower. **(b) Risk aversion in bees evolved in a risk-less environment** - Preference for blue flowers in 100 test trials averaged over 40 bees evolved in a risk-less environment which contained two constantly-rewarding flower types. As in (a), risk averse behavior is prominent, although the two flower types yield the same mean reward. **(c) Risk aversion is ordered by learning rate.** Each point represents the percentage of visits to constant-rewarding flowers in 50 test trials averaged over 40 previously evolved bees, tested with a clamped learning rate.

Figure 6

Probability Matching - Preference for blue flowers in 100 test trials averaged over 40 previously evolved bees, now tested in probability matching conditions. All flowers yield $1\mu l$ nectar but with different reward probabilities. Reward probabilities for blue and yellow flowers respectively were (a) 0.8,0.4 (b) 0.8,0.2 (contingencies switched at trial 50). Horizontal lines - behavior predicted by perfect matching.

Figure 7

Synaptic weights of a mobile robot incorporating a NN controller of one of the previously evolved bees, performing 20 foraging trials (blue flowers - $1/2\mu l$ nectar, yellow - $1\mu l$ in half the flowers, contingencies switched after trial 10). **(a)** The foraging robot. **(b)** Blue and yellow weights in the differential module represent the rewards expected from the two flower colors along the trials. **Top:** Flower color chosen in each trial.

Table 1

Heterosynaptic learning rule coefficients in five successful evolutionary runs. Columns 1-4: Mean(standard deviation) of the evolved learning rule coefficients of the differential module heterosynaptic learning rule, in the last generation of five successful runs. Columns 5-8: Number of bees in the last generation with a positive value of B, C, D and $(D - B)$ respectively.

References

Ackley, D., & Littman, M. (1991). Interactions between learning and evolution. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial life II.* Addison-Wesley.

Bailey, C., Giustetto, M., Huang, Y., Hawkins, R., & Kandel, E. (2000). Is heterosynaptic modulation esential for stabilizing hebbian plasticity and memory? *Nature Reviews Neuroscience*, *1*, 11-20.

Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 215-232). Cambridge: MIT Press.

Baxter, J. (1992). The evolution of learning algorithms for artificial neural networks. In D. Green & T. Bossomaier (Eds.), *Complex systems.* IOS Press.

Breiman, L. (1968). *Probability.* Addison-Wesley.

Chalmers, D. J. (1990). The evolution of learning: An experiment in genetic connectionism. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.), *Proc. of the 1990 connectionist models summer school.* Morgan Kaufmann.

Dittman, J. S., & Regehr, W. G. (1997). Mechanism and kinetics of heterosynaptic depression at a cerebellar synapse. *The journal of neuroscience*, *17*(23), 9048-9059.

Donahoe, J. W., & Packard-Dorsel, V. (Eds.). (1997). *Neural network models of cognition: Biobehavioral foundations.* Elsevier Science.

Fellous, J.-M., & Linster, C. (1998). Computational models of neuromodulation: A review. *Neural Computation*, *10*, 791-825.

Floreano, D., & Mondada, F. (1996). Evolution of homing navigation in a real mobile robot. *IEEE Transactions on Systems, Man and Cybernetics*, *26*(3), 396-407.

Floreano, D., & Mondada, F. (1998). Evolutionary neurocontrollers for autonomous mobile robots. *Neural networks*, *11*, 1461-1478.

Fontanari, J. F., & Meir, R. (1991). Evolving a learning algorithm for the binary perceptron. *Network*, *2*(4), 353-359.

Graybiel, A. M., & Kimura, M. (1995). Adaptive neural networks in the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 103-116). Cambridge: MIT Press.

Greggers, U., & Menzel, R. (1993). Memory dynamics and foraging strategies of honeybees. *Behavioral Ecology and Sociobiology*, *32*, 17-29.

Hammer, M. (1993). An identified neuron mediates the unconditioned stimulus in associative learning in honeybees. *Nature*, *366*, 59-63.

Hammer, M. (1997). The neural basis of associative reward learning in honeybees. *Trends in Neuroscience*, *20*(6), 245-252.

Harder, L. D., & Real, L. A. (1987). Why are bumble bees risk averse? *Ecology*, *68*(4), 1104-1108.

Hardy, G. H., Littlewood, J. E., & Polya, G. (1934). *Inequalities.* Cambridge University Press.

Hinton, G. E., & Nowlan, S. J. (1987). How learning guides evolution. *Complex Systems*, *1*, 495-502.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 249-270). Cambridge: MIT Press.

Houk, J. C., & Wise, S. P. (1995). Distributed modular architecture linking basal ganglia, cerebellum and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex*, *5*, 95-111.

Kacelnik, A., & Bateson, M. (1996). Risky thoeries - the effect of variance on foraging decisions. *American Zoologist*, *36*, 402-434.

Kaesar, T., Rashkovich, E., Cohen, D., & Shmida, A. (Submitted). Choice behavior of bees in two-armed bandit situations: Experiments and possible decision rules. *Behavioral Ecology.*

Kimura, M., & Matsumoto, N. (1997). Nigrostriatal dopamine system may contribute to behavioral learning through providing reinforcement signals to the striatum. *European Neurology*, *38*(suppl 1), 11-17.

March, J. G. (1996). Learning to be risk averse. *Psychological Review*, *103*(2), 309-319.

Menzel, R., & Muller, U. (1996). Learning and memory in honeybees: From behavior to neural substrates. *Annual Reviews in Neuroscience*, *19*, 379-404.

Mitchell, T. (1997). *Machine learning.* McGraw Hill.

Montague, P. R. (1997). Biological substrates of predictive mechanisms in learning and action choice. In J. W. Donahoe & V. Packard-Dorsel (Eds.), *Neural network models of cognition: Biobehavioral foundations* (p. 406-421). Elsevier Science.

Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., & Sejnowski, T. J. (1993). Using aperiodic reinforcement for directed self-organization. In C. L. Giles, S. J. Hanson, & J. D. Cowan (Eds.), *Advances in neural information processing systems* (Vol. 5, p. 969-976). Morgan Kaufmann.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*, 725-728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framewok for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience*, *16*, 1936-1947.

Nolfi, S., Elman, J. L., & Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behavior*, *3*(1), 5-28.

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, *253*, 980-985.

Real, L. A. (1996). Paradox, performance and the architecture of decision making in animals. *American Zoologist*, *36*, 518-529.

Rothschild, M., & Stiglitz, J. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, *2*, 225-243.

Schacher, S., Wu, F., & Sun, Z. Y. (1997). Patway-specific synaptic plasticity: activity-dependent enhancement and suppression of long-term heterosynaptic facilitation at converging inputs on a single target. *The Journal of Neuroscience*, *17*(2), 597-606.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1-27.

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, *1*, 199-207.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1599.

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 233-248). MIT Press.

Seth, A. K. (1999). Evolving behavioral choice: An investigation into Herrnstein's matching law. In D. Floreano, J. Nicoud, & F. Mondada (Eds.), *Advances in artificial life, 5th European conference, ECAL '99* (p. 225-235). Lausanne, Switzerland: Springer.

Smallwood, P. D. (1996). An introduction to risk sensitivity: The use of Jensen's inequality to clarify evolutionary arguments of adaptation and constraint. *American Zoologist*, *36*, 392-401.

Sutton, R. S. (1988). Learning to predict by the method of temporal difference. *Machine Learning*, *3*, 9-44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* MIT Press.

Suzuki, T., Miura, M., Mishimura, K., & Aosaki, T. (2001). Dopamine-dependent synaptic plasticity in the striatal cholinergic interneurons. *The Journal of Neuroscience*, *21*(17), 6492-6501.

Thuijsman, F., Peleg, B., Amitai, M., & Shmida, A. (1995). Automata, matching and foraging behavior of bees. *Journal of Theoretical Biology*, *175*, 305-316.

Unemi, T., Nagayoshi., M., Hirayama, N., Nade, T., Yano, K., & Masujima, Y. (1994). Evolutionary differentiation of learning abilities - a case study on optimizing parameter values in Q-learning by a genetic algorithm. In R. A. Brooks & P. Maes (Eds.), *Artificial life IV* (p. 331-336). MIT Press.

Vogt, K. E., & Nicoll, R. E. (1999). Glutamate and gama-aminobutyric acid mediate a heterosynaptic depression at mossy fiber synapses in the hippocampus. *Proceedings of the national academy of science, USA*, *96*, 1118-1122.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43-48.

Wickens, J., & Kötter, R. (1995). Cellular models of reinforcememnt. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 187-214). MIT Press.

Wilson, S. W. (1996). Explore/exploit strategies in autonomy. In P. Maes, M. Mataric, J. Pollack, J. A. Meyer, & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the fourth international conference on the simulation of adaptive behavior.* Cambridge, MA: The MIT Press/Bradford Books.

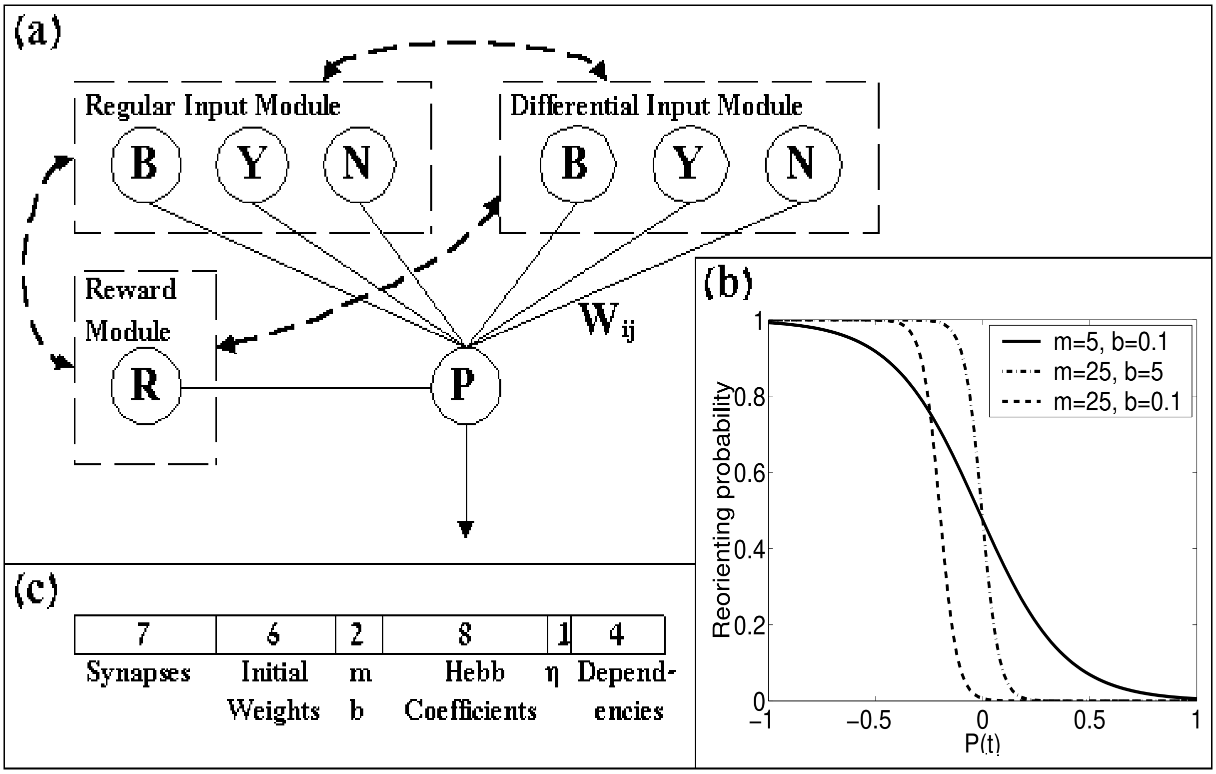
(a)
Regular Input Module
Differential Input Module
B
Y
N
B
Y
N
Reward Module
R
P
W_{ij}
(b)
m=5, b=0.1
m=25, b=5
m=25, b=0.1
Reorienting probability
P(t)
(c)
7
6
2
8
1
4
Synapses
Initial Weights
m b
Hebb Coefficients
η
Dependencies

Figure 1

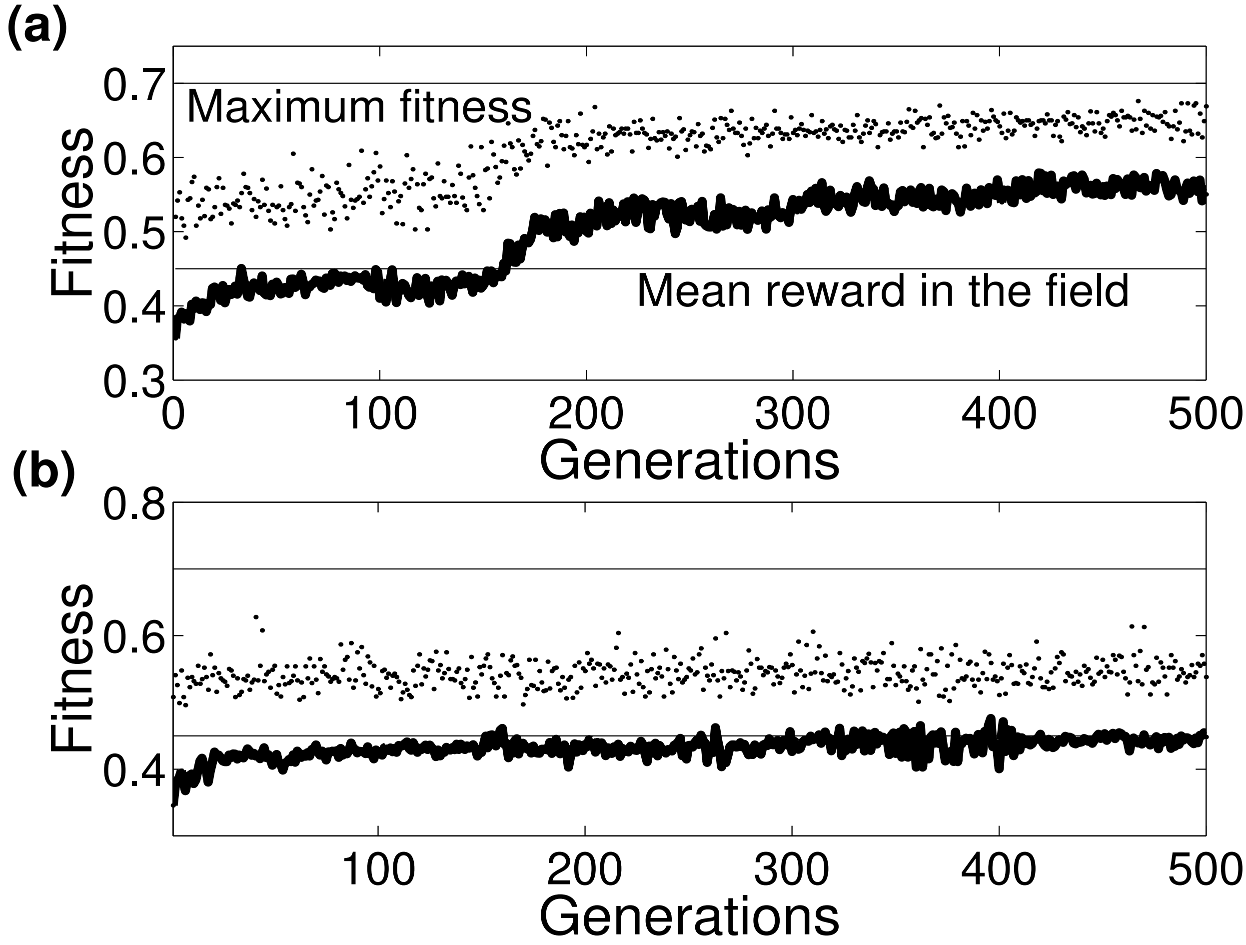
(a)
0.7
0.6
0.5
0.4
0.3
Fitness
Maximum fitness
Mean reward in the field
0
100
200
300
400
500
Generations
(b)
0.8
0.6
0.4
Fitness
100
200
300
400
500
Generations

Figure 2

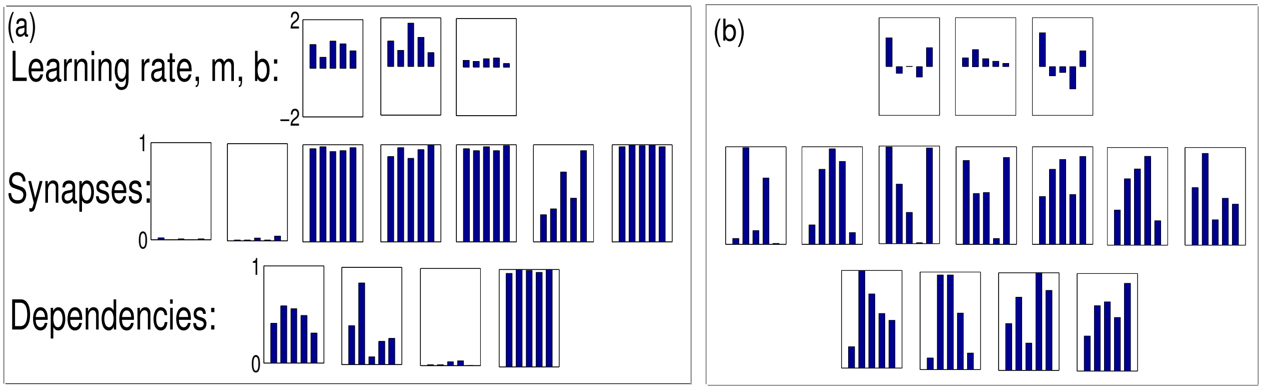

Figure 3

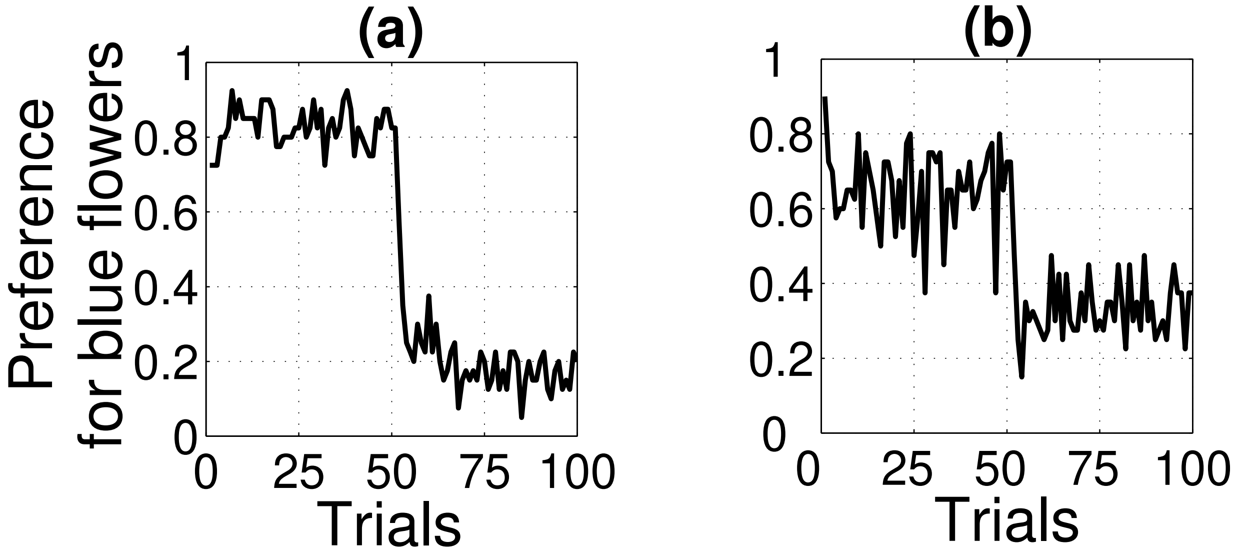
(a)
(b)
Preference
for blue flowers
1
0.8
0.6
0.4
0.2
0
0
25
50
75
100
Trials
1
0.8
0.6
0.4
0.2
0
0
25
50
75
100
Trials

Figure 4

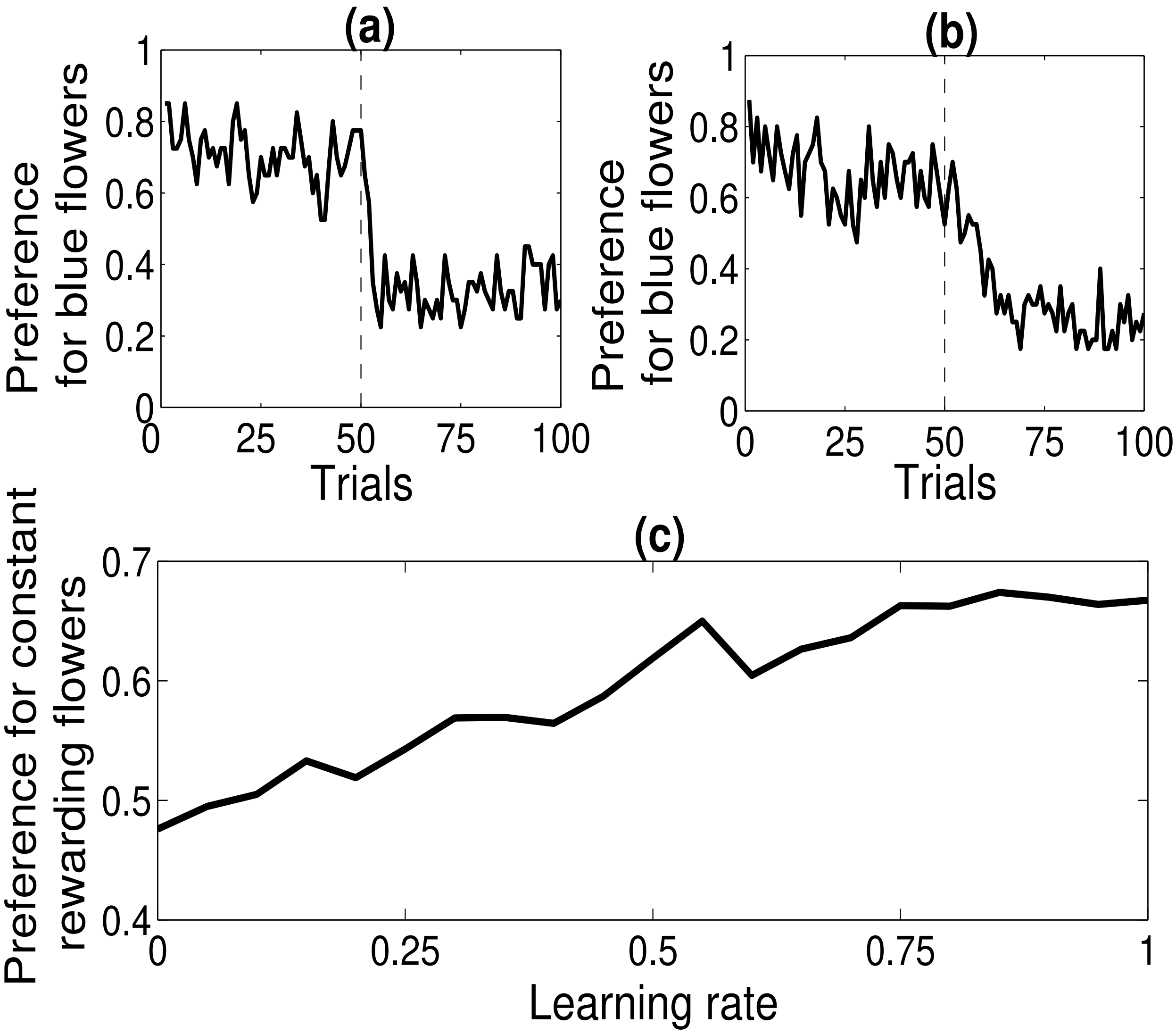
(a)
Preference
for blue flowers
1
0.8
0.6
0.4
0.2
0
0
25
50
75
100
Trials
(b)
Preference
for blue flowers
1
0.8
0.6
0.4
0.2
0
0
25
50
75
100
Trials
(c)
Preference for constant
rewarding flowers
0.7
0.6
0.5
0.4
0
0.25
0.5
0.75
1
Learning rate

Figure 5

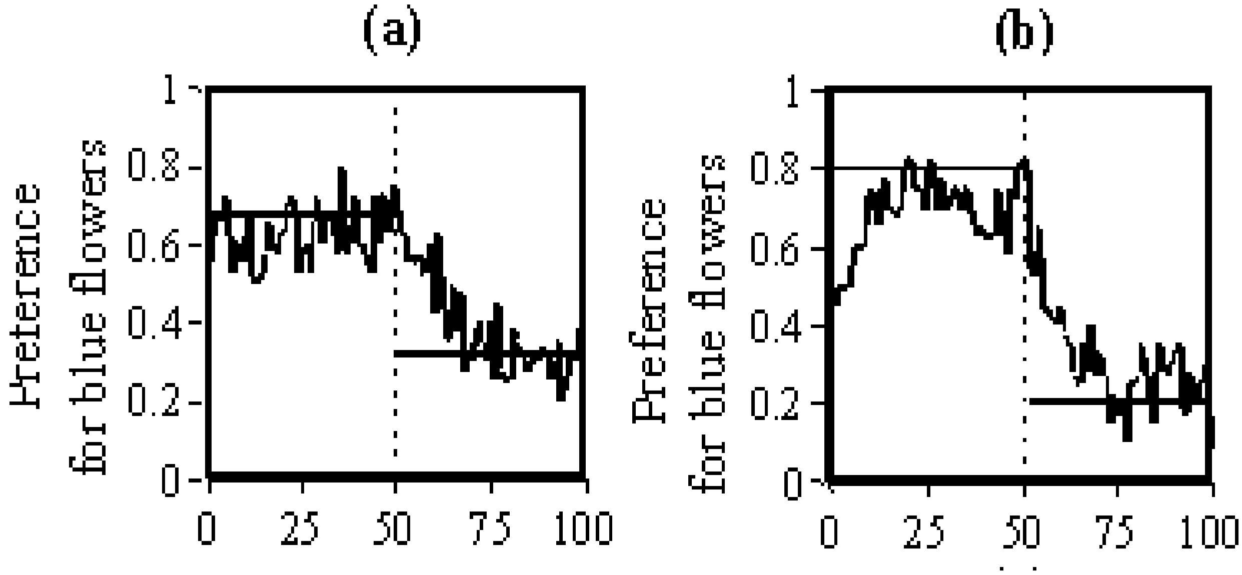

Figure 6

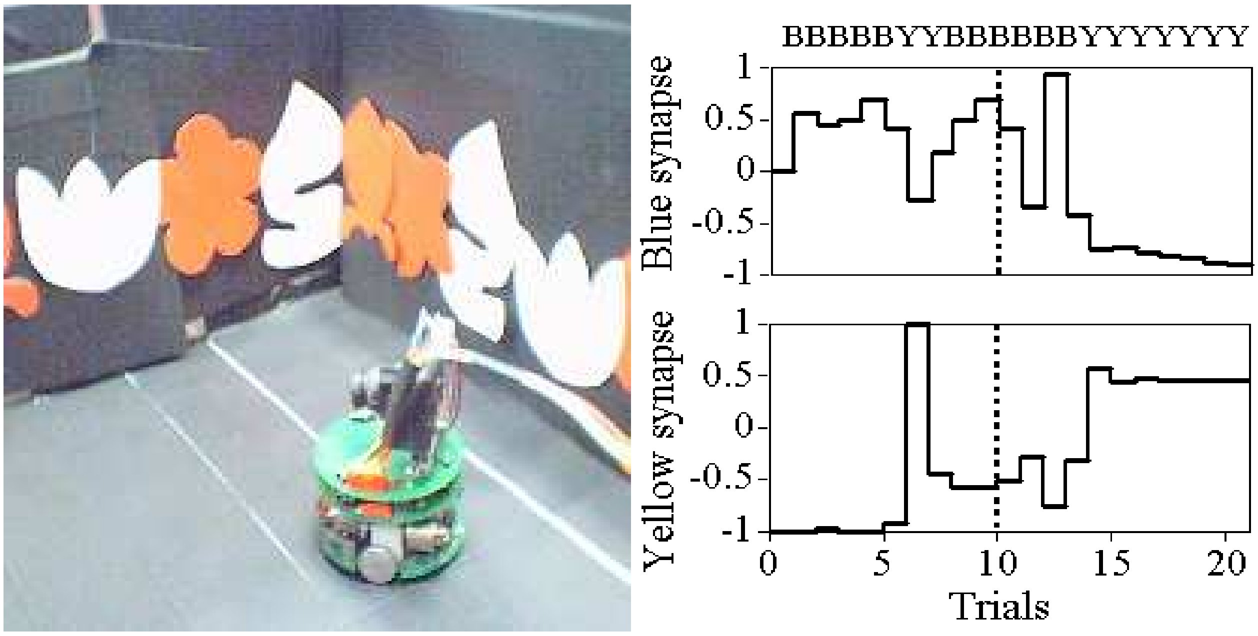
BBBBBYYBBBBBBYYYYYYY
Blue synapse
Yellow synapse
1
0.5
0
-0.5
-1
0
5
10
15
20
Trials

Figure 7

Run	A	B	C	D	(B>0)	(C>0)	(D>0)	(D-B>0)
1	-1.2(0.1)	0.1(0.1)	-0.1(0.1)	-0.1(0.1)	73	24	1	15
2	-1.2(0.2)	-0.2(0.1)	0.2(0.1)	-0.1(0.2)	0	100	47	61
3	-1.7(0.2)	0.3(0.1)	-0.3(0.1)	0.0(0.1)	94	0	82	7
4	-1.3(0.1)	0.2(0.1)	0.3(0.1)	-0.2(0.1)	98	100	1	0
5	-1.0(0.1)	0.0(0.3)	0.3(0.1)	-0.5(0.1)	66	100	0	2
Mean	-1.3(0.3)	0.1(0.2)	0.1(0.3)	-0.2(0.2)	66	65	26	17

Table 1