Evolution of Reinforcement Learning in Uncertain Environments: Emergence of Risk-Aversion and Matching

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

[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, ruppin@math.tau.ac.il

Abstract. Reinforcement learning (RL) is a fundamental process by which organisms learn to achieve a goal from interactions with the environment. We use Artificial Life techniques to derive (near-)optimal neuronal learning rules in a simple neural network model of decision-making in simulated bumblebees foraging for nectar. The resulting networks exhibit efficient RL, allowing the bees to respond rapidly to changes in reward contingencies. Furthermore, the evolved synaptic plasticity dynamics give rise to varying exploration/exploitation levels from which emerge the well-documented foraging strategies of risk aversion and probability matching. These are shown to be a direct result of optimal RL, providing a biologically founded, parsimonious and novel explanation for these behaviors. Our results are corroborated by a rigorous mathematical analysis and by experiments in mobile robots.

1 Introduction

Reinforcement learning (RL) is a process by which organisms learn from their interactions with the environment to achieve a goal [14]. In RL, learning is contingent upon a scalar reinforcement signal which only provides evaluative information about how good an action is in a certain 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. A computational understanding of neuronal reinforcement learning is a necessary step towards an understanding of brain functions, 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 learning rules for maximizing reward in RL?** In this paper, we use Artificial-life (Alife) techniques to derive the **optimal neuronal learning**

rules that give rise to efficient RL in uncertain environments. We further investigate the behavioral strategies which emerge from optimal RL.

RL has been demonstrated and studied extensively in foraging bees. Real [2] 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, 85% of the bees' visits were to the blue constant-rewarding flowers, although the mean return from both flower types was identical. Such risk-averse behavior has also been demonstrated elsewhere [1], and has traditionally been accounted for by hypothesizing the existence of a subjective non-linear concave "utility function" for nectar [6]. Risk averse behavior is also prominent in humans, and is **an important choice strategy, well-studied in economics and game-theory**, although its biological basis is not yet firmly established.

A foraging bee deals with a rapidly changing environment - parameters such as the weather, and competition affect the availability of rewards from different kinds of flowers. This implies an "armed-bandit" type scenario, in which the bee collects food and information simultaneously. As a result there exists a tradeoff between exploitation and exploration, as the bee's actions directly effect the "training examples" which it will encounter through the learning process. A notable strategy by which bumblebees (and other animals) optimize choice in such situations is probability matching. When faced with flowers offering similar rewards but with different probabilities, bees match their choice behavior to the reward probabilities of the flowers [16]. This seemingly "irrational" behavior with respect to optimization of reward intake is explained as an Evolutionary Stable Strategy (ESS) for the individual forager when faced with competitors [10], as it produces an Ideal Free Distribution (IFD) in which the average intake of food is the same at all food sources. Using Alife techniques, Seth evolved battery-driven agents competing for two different battery refill sources, and showed that indeed matching behavior emerges only in a multi-agent scenario [3].

In a previous neural network (NN) model, Montague et al. [13] simulated bee foraging in a 3D arena of blue and yellow flowers, based on a neurocontroller modelled after an identified interneuron in the honeybee suboesophogeal ganglion. This neuron's activity represents the reward value of gustatory stimuli, and similar to Dopaminergic neurons in the Basal Ganglia, is activated by unpredicted rewards [12]. 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 flying in the same direction, or to change direction randomly. Upon landing, a reward is received according to the subjective utility of the nectar content of the chosen flower [6], and the synaptic weights of the networks are updated according to a special anti-Hebbian-like learning rule. As a result, the values of the weights come to represent the expected rewards from each flower type.

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. In contrast to common Alife applications which involve NNs with evolvable synaptic weights or architectures [4, 5, 15], 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 [7, 9]. We define a general framework for evolving learning rules, which essentially encompasses **all heterosynaptic Hebbian learning rules**, along with other characteristics of the learning dynamics. 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.

In the following section we describe the model and the evolutionary dynamics. Section 3 describes the results of our simulations, and the evolution of RL. In section 4 we analyze the foraging behavior resulting from the learning dynamics, and find that when tested in new environments, our Alife creatures manifest risk aversion and probability matching behaviors. Although this behavior was not selected for, we rigorously prove that these strategies emerge directly from optimal RL. Section 5 describes a minirobot implementation of the evolved RL model, and we conclude with a discussion of the results in section 6.

2 The Model

A simulated bee-agent flies above a 3D patch of 60x60 randomly scattered blue and yellow flowers. A bee's life consists of 100 trials. In each trial the bee starts its descent from a height of 10 units, 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 bee consumes any available nectar in one timestep, and another trial begins. **The evolutionary goal (the fitness criterion) is to maximize nectar intake.**

In the neural network controlling the bee's flight (Fig. 1a), which is an extension of Montague et al's network [13], 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

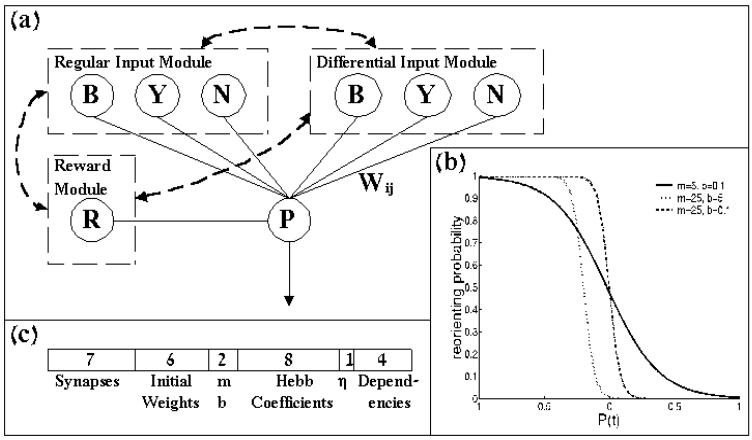

Fig. 1. (a) The bee's neural network controller. (b) The bee's action function. Probability of reorienting direction of flight as a function of $P(t)$ for different values of parameters m, b. **(c) The genome sequence of the simulated bee.**

there is no new input $[X_i(t) = 0]$), and zero during flight. Note that **we do not incorporate any form of utility function** with respect to the reward. Thus P's continuous-valued output is:

$$P(t) = R(t) + \sum_{i \in \text{regular}} W_i X_i(t) + \sum_{i \in \text{differential}} W_i[X_i(t) - X_i(t-1)]. \quad (1)$$

The bee's action is determined according to the output $P(t)$ using Montague et al's probabilistic action function [6, 13] (Fig. 1b):

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

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

$$\Delta W_i = \eta(AX_i(t)P(t) + BX_i(t) + CP(t) + D), \quad (3)$$

where η is a global learning rate parameter, $X_i(t)$ and $P(t)$ are the presynaptic and the postsynaptic 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 Fig. 1a), such that if module Z depends on module Y, Z's synaptic weights will be updated according to equation (3) only if module Y's respective neurons have fired (if it is not dependent, the weights will be updated on every timestep). Thus the bee's "brain" is capable of **a non-trivial axo-axonic gating of synaptic plasticity.**

The simulated bee's genome (Fig. 1c) consists of a string of 28 genes, each representing a parameter governing the network architecture and or its learning dynamics. Seven boolean genes determine whether each synapse in the network

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 synapses); 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: The regular and differential modules each have a learning rule specified by 4 real-valued genes (parameters A-D of equation (3), range [-1,1]); The global learning rate 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 (with **no Lamarkian inheritence** of learned weights) after performing recombination (genewise, $p = 0.25$) 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. One hundred offsprings were created, these once again tested in the flower field. This process continued for a large number of generations.

3 Evolution of Reinforcement Learning

To promote the evolution of efficient learning rules, 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: runs in which reward-dependent choice behavior is successfully evolved are characterized by two distinct evolutionary jumps (Fig. 2a), while unsuccessful runs (which produce behavior that is not dependent on rewards) show only the first jump.

About half of the evolutionary runs were successful. Figure 2b 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. All successful networks have a specific architecture which includes the reward, differential blue and differential yellow synapses, as well as a dependency of the differential module on the reward module, conditioning modification of these synapses on the presence of reward. Thus we find that in our framework, only a network architecture similar to that used by Montague et

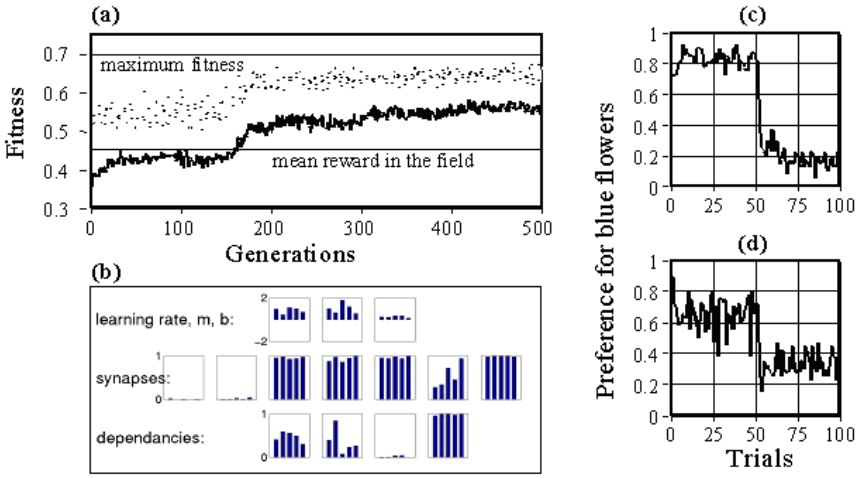

Fig. 2. (a) Typical fitness scores of a successful run of 500 generations. Solid line - mean fitness, dotted line - maximum fitness in each generation. **(b) Mean value of several genes** in the last generation of five successful runs. Each subfigure shows the mean value of one gene in the last generation of five runs. **(c,d) Preference for blue flowers for two different bees** from the last generation of a successful run, averaged over 40 test bouts, each consisting of 100 trials. Blue is the initial constant-rewarding high-mean flower. Reward contingencies are switched at trial 50.

al. [13] can produce above-random foraging behavior, supporting their choice as an optimal one. However, **our optimized networks utilize a heterosynaptic learning rule different from that used by Montague et al., which gives rise to several important behavioral strategies.**

Bees from the last generation of a successful run show 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. An examination of the behavior of the evolved bees, reveals that there are individual differences between the bees in their degree of exploitation of the high-rewarding flowers versus exploration of the other flowers (Fig. 2c,d). This can be explained by an **interesting relationship between the micro-level Hebb rule coefficients and the exploration/exploitation tradeoff characteristic of the macro-level behavior:** In the common case when upon landing the bee sees only one color, the synaptic update rule for the corresponding differential synapse is

$$\Delta W_i(t+1) = \eta[(A-C)\cdot(-1)\cdot[R(t)-W_i(t)]+(D-B)] \tag{4}$$

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

$$\Delta W_j(t+1) = \eta(C\cdot[R(t)-W_i(t)]+D). \tag{5}$$

Thus, positive C and D values result in spontaneous strengthening of competing synapses, leading to an exploration-inclined bee. Negative values will result in a declining tendency to visit competing flower types, leading to exploitation-inclined behavior.

4 Emergence of Risk Aversion and Probability Matching

A prominent strategy exhibited by the evolved bees is risk-aversion. Figure 3a shows the choice behavior of previously evolved bees, tested in a new environment where the mean rewards of the two kinds of flowers are identical. Although the situation does not call for any flower preference, the bees prefer the constant-rewarding flower. Furthermore, bees evolved in an environment containing two constant-rewarding flowers yielding different rewards, also exhibit risk-averse behavior when tested in a variable-rewarding flower scenario, thus risk- aversion is not a consequence of evolution in an uncertain environment per se. In contradistinction to the conventional explanations of risk aversion, our model does not include a non-linear utility function. **What hence brings about risk-averse behavior in our model?** Corroborating previous numerical results [11], we prove analytically that this foraging strategy is a direct consequence of Hebbian learning dynamics in an armed-bandit-like RL situation.

The bee's stochastic foraging decisions can be formally modeled as choices between a variable-rewarding (v) and a constant-rewarding (c) flower, based on memory (synaptic weights). We consider the bee's long-term choice dynamics as a sequence of N cycles, each choice of (v) beginning a cycle. The frequency f_v of visits to (v) can be determined (via Birkhoff's Ergodic theorem) by the expected number of visits to (c) in a cycle, and is

$$f_v = \frac{1}{E[1/p_v(W_v, W_c)]} \tag{6}$$

where $p_v(W_v, W_c)$ is the probability of choosing (v) in a trial in which the synaptic weights are W_v and W_c for the variable and the constant flower respectively. We show that if $p_v(W_v, W_c)$ is a positive increasing choice function such that $[1/p_v(W_v, W_c)]$ is convex, 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 this analytical result (Fig. 3b).

In essence, due to 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 ($\sim$0.8). After landing on an empty flower of the variable- rewarding type, the bee updates the reward expectation to near zero, and as a result, prefers the constantly rewarding flower, from which it constantly expects (and recieves) 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

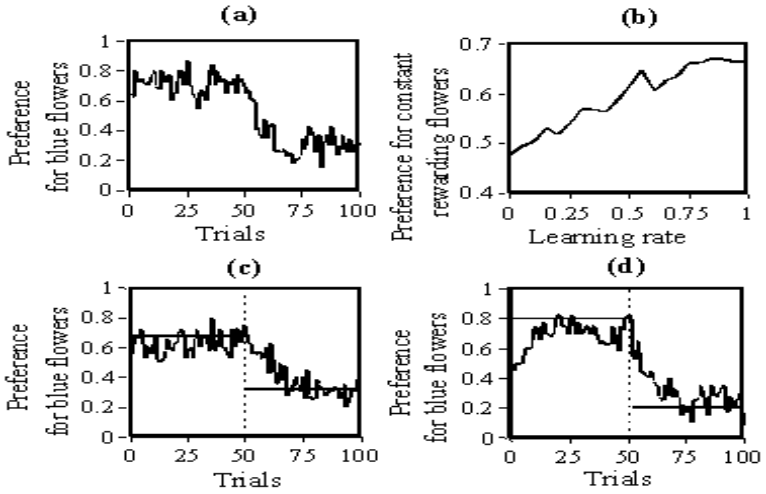

Fig. 3. Preference for blue flowers averaged over 40 previously evolved bees tested in conditions different from those they were evolved in: **(a) Risk aversion** - Although both flower types yield the same mean reward (blue - $\frac{1}{2}\mu l$ nectar, yellow - $1\mu l$ in half the flowers, contingencies switched at trial 50), there is a marked preference for the constant-yielding flower. **(b) Risk aversion is ordered according to learning rate.** Each point represents the percentage of visits to constant-rewarding flowers in 50 test trials averaged over 40 previously evolved bees, with a clamped learning rate. **(c-d) Matching** - All flowers yield $1\mu l$ nectar with reward probabilities for blue and yellow flowers (c) 0.8, 0.4 and (d) 0.8, 0.2 respectively (contingencies switched at trial 50). Horizontal lines - behavior predicted by perfect matching.

variable flower yields a high reward, the preference for this flower will be short lived, lasting only until the next unrewarded visit. Note that such abnormally high learning rates were also used in Montague et al.'s [13] model, and have been hypothesized by Real [2].

The simulated bees also demonstrate probability-matching behavior. Figure 3(c,d) 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 is again a direct result of the learning dynamics Thus, in contradistinction to previous accounts, matching can be evolved in a non-competitive setting, as a direct consequence of optimal RL.

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. 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 RL and risk-averse behavior (Fig. 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.

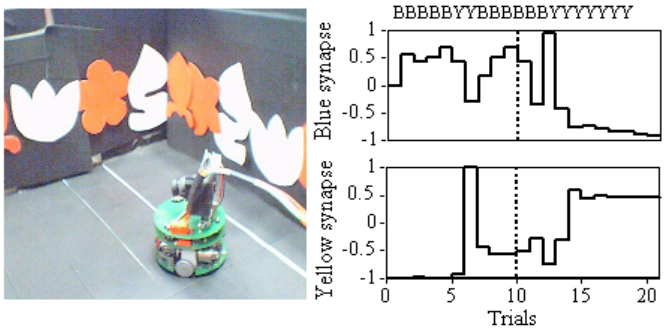

Fig. 4. Synaptic weights of a mobile robot incorporating a NN controller of one of the previously evolved bees, performing 20 foraging trials (blue flowers - $\frac{1}{2}\mu l$ nectar, yellow - $1\mu l$ in half the flowers, contingencies switched after trial 10). **(a)** The foraging robot. **(b)** Blue and yellow differential weights represent the expected rewards from the two flower colors along the trials. **Top:** Flower color chosen in each trial.

6 Discussion

The interplay between learning and evolution has been previously investigated in the field of Alife. Much of this research has been directed to elucidating the relationship between evolving traits (such as synaptic weights) versus learning them [4, 8]. 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 [7, 5]. 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 heterosynaptic learning rules are by no means trivial, as they include an anti-Hebbian monosynaptic term and employ axo-axonic plasticity modulation. We have no rigorous proof as to their optimality, but results from multiple evolutionary runs strongly suggest this.

The emergence of complex foraging behaviors as a result of optimal learning per se, demonstrate once again the strength of Alife as a methodology that links together phenomena on the neuronal and the behavioral levels. We show

that the fundamental macro-level strategies of risk aversion and matching are a direct result of the micro level synaptic learning dynamics, which control the tradeoff between exploration and exploitation making additional assumptions conventionally used to explain them redundant. This result is important not only to the fields of Alife and animal learning theories, but also to economics and game theory.

In summary, the significance of this work is two-fold: on one hand we show the strength of simple Alife models in evolving fundamental processes such as reinforcement learning, and on the other we show that optimal reinforcement learning can directly explain complex behaviors such as risk aversion and probability matching, without need for further assumptions.

References

[1] Kacelnik A. and Bateson M. Risky thoeries - the effect of variance on foraging decisions. *American Zoologist*, 36:402–434, 1996.

[2] Real L. A. Animal choice behavior and the evolution of cognitive architecture. *Science*, 253:980–985, August 1991.

[3] Seth A.K. Evolving behavioral choice:an investigation into herrnstein's matching law. In Floreano D., Nicoud J., and Mondada F., editors, *Advances in Artificial Life, 5th European Conf., ECAL '99*, pages 225–235. Springer, 1999.

[4] Ackley D. and Littman M. Interactions between learning and evolution. In Langton C.G., Taylor C., Farmer J. D., and Rasmussen S., editors, *Artificial Life II*. Addison-Wesley, 1991.

[5] Floreano D. and Mondada F. Evolution of homing navigation in a real mobile robot. *IEEE Trans. on Systems, Man and Cybernetics*, 26(3):396–407, 1996.

[6] Harder L. D. and Real L. A. Why are bumble bees risk averse? *Ecology*, 68(4):1104–1108, 1987.

[7] Chalmers D.J. The evolution of learning: An experiment in genetic connectionism. In Touretzky D.S., Elman J.L., Sejnowski T.J., and Hinton G.E., editors, *Proc. of the 1990 Connectionist Models Summer School*. Mogan Kaufmann, 1990.

[8] Hinton G. E. and Nowlan S. J. How learning guides evolution. *Complex Systems*, 1:495–502, 1987.

[9] Fontanari J. F. and Meir R. Evolving a learning algorithm for the binary perceptron. *Network*, 2(4):353–359, November 1991.

[10] Thuijsman F., Peleg B., Amitai M., and Shmida A. Automata, matching and foraging behavior of bees. *Journal of Theoretical Biology*, 175:305–316, 1995.

[11] March J. G. Learning to be risk averse. *Psych. Review*, 103(2):309–319, 1996.

[12] Hammer M. The neural basis of associative reward learning in honeybees. *Trends in Neuroscience*, 20(6):245–252, 1997.

[13] Montague P.R., Dayan P., Person C., and Sejnowski T.J. Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, 377:725–728, 1995.

[14] Sutton R.S. and Barto A.G. *Reinforcement learning:An introduction*. MIT Press, 1998.

[15] Nolfi S., Elman J. L., and Parisi D. Learning and evolution in neural networks. *Adaptive Behavior*, 3(1):5–28, 1994.

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