TEL AVIV UNIVERSITY
Department of Psychology

Evolution of Reinforcement Learning in Uncertain Environments

Thesis submitted as part of the requirements for the degree of Master of Arts in Psychology

by

Yael Niv

Supervised by:
Dr. Daphna Joel
and
Prof. Eytan Ruppin

March 2001

Acknowledgements

I am grateful to Daphna Joel and Eytan Ruppin, my supervisors, for their invaluable guidance and support. To Daphy for boldly venturing into the field of neural network modelling and contributing her clear thoughts, sharp distinctions and original ideas to this work. To Eytan for his wisdom, his endless enthusiasm and his sincere appreciation of the significance of my work.

Many thanks to Prof. Isaac Meilijson for the mathematical proof regarding the emergence of risk-aversion.

I thank Dr. Tamar Keasar for introducing me to the BeeHave lab at HUJI, for providing me with the probability matching data, and for her invaluable comments and new ideas throughout my research.

I have benefitted from discussions with the many people who have read drafts of this work or listened to my talks. All these have helped me immensely in making my ideas coherent and understandable.

Special thanks to my family, and especially my father, Yehuda, who followed my research closely, nagged until I finally sat down to write this, and helped me tackle the difficult parts.

Abstract

Reinforcement learning is a fundamental process by which organisms learn to achieve a goal from interactions with the environment. Using Artificial Life 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 from which emerge the well-documented choice strategies of risk aversion and probability matching. These strategies are shown to be a direct result of reinforcement learning, providing a biologically founded, parsimonious and novel explanation for these behaviors. Our results are corroborated by a rigorous mathematical analysis and their robustness in real-world situations is supported by experiments in a mobile robot.

1 Introduction

Reinforcement learning (RL) is a process by which organisms learn from their interactions with the environment to achieve a goal [28]. 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 [5]. A computational understanding of neuronal reinforcement learning is a necessary step towards an 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 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 reinforcement learning in uncertain environments.** We further investigate the behavioral strategies, which emerge as a result of optimal RL.

RL has been demonstrated and studied extensively in foraging bees. Real [23, 24] 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 [15]: 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-averse behavior has also been demonstrated in other animals (see [15] for a review), and has traditionally been accounted for by hypothesizing the existence of a non-linear concave "utility function" for reward. 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 [12, 27].

A foraging bee deals with a rapidly changing environment - parameters such as the weather, the season, and competition all 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 [9]. As a result there exists a tradeoff between exploitation and exploration, as the bee's actions directly affect the "training examples" which it will encounter through the trial-and- error learning process. Such a tradeoff is demonstrated in Real's experiment [23] by the $\sim 15\%$ sampling of the variable flower which ensures the bee that changes in the reward contingencies in that flower will be detected.

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 choose the different flower types by a ratio that matches the reward ratio of the flowers [9, 16]. This seemingly "irrational" behavior with respect to optimization reward intake is explained as an Evolutionary Stable Strategy (ESS) for the individual forager, when faced with competitors [29]. In a multi-animal competitive setting, matching strategy 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 Alife techniques, Seth [26] 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.

In a previous neural network (NN) model, Montague et al. [21] 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 [10]. This neuron's activity represents the reward value of gustatory stimuli, and similar to Dopaminergic neurons in the Basal Ganglia of primates, is activated by unpredicted rewards and by reward predicting stimuli, and is not activated by predicted rewards [11]. 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 is received according to the subjective utility of the nectar content of the chosen flower [12], 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 [20]. 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 [19]. In contrast to common Alife applications which involve NNs with evolvable synaptic weights or architectures [1, 6, 22], 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 [2, 4, 7, 8], or evolved only a subset of the learning rule parameters [30]. 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, such as learning dependencies between modules. 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 reports the results of our simulations: In 3.1 we describe the successful evolution of RL, and the conditions which need to be met to facilitate such evolution. Section 3.2 describes the evolved synaptic update rule, and its influence on the exploration/exploitation tradeoff of the foraging bees. In section 3.3 we further analyze the foraging behaviors resulting from the learning dynamics, and find that when tested in new environments, our evolved bees manifest risk aversion and probability matching behaviors. Although these choice strategies were not selected for, we rigorously prove that risk-aversion and probability matching emerge directly from optimal RL. Section 3.4 describes a minirobot implementation of the model, aimed at assessing its robustness. We conclude with a discussion of the results in section 4.

2 Methods

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 ~ 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" 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.**

In the neural network controlling the bee's flight (Figure 1a), which is an extension of Montague et al's network [21], 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

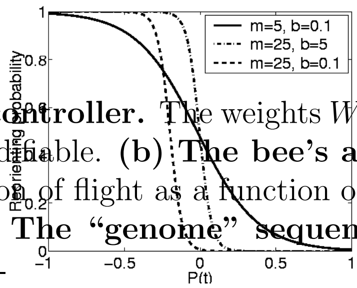
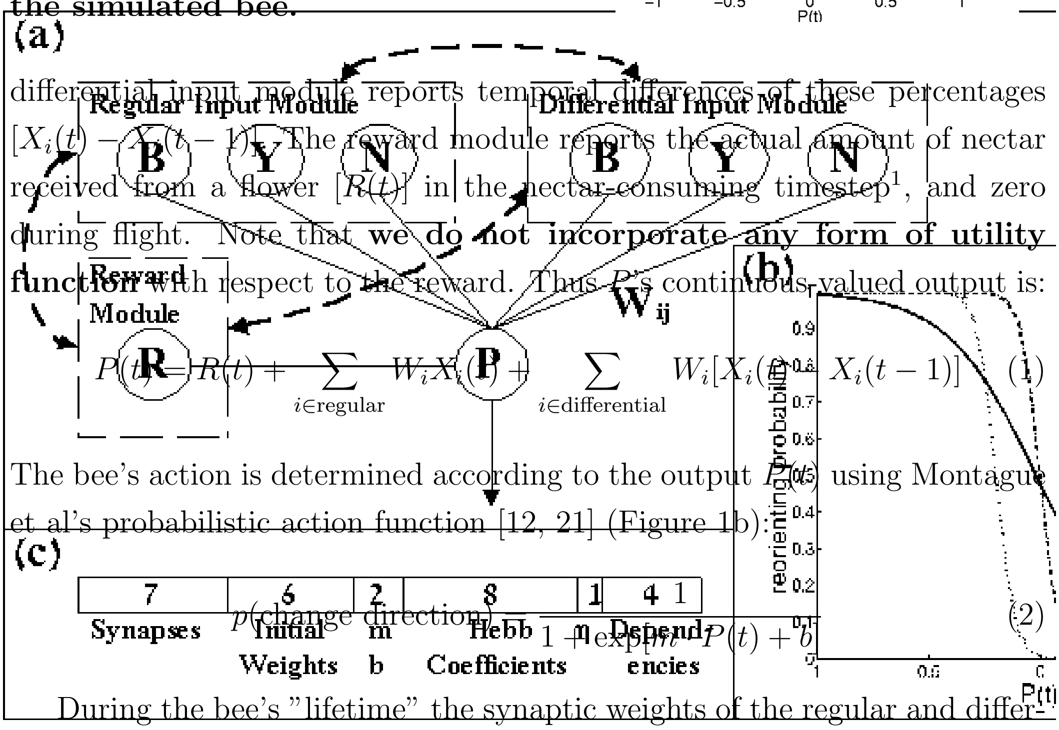


Figure 1: **(a) The bee's neural network controller.** The weights $W_i(t)$ of the regular and differential modules are modifiable. **(b) The bee's action function.** Probability of reorienting direction of flight as a function of $P(t)$ for different values of parameters m, b. **(c) The "genome" sequence of the simulated bee.**



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[1], 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 [12, 21] (Figure 1b):

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

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

[1] In this timestep it is also assumed that there is no new input $[X_i(t) = 0]$.

form:

$$\Delta W_i(t) = \eta[AX_i(t)P(t) + BX_i(t) + CP(t) + D] \tag{3}$$

where η is a global learning rate parameter, $X_i(t)$ 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 1a), 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[2], and 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 (Figure 1c) consists of a string of 28 genes, each representing a parameter governing the network architecture and or its 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[3]; 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

[2]A dependency on the reward module is satisfied when the reward neuron fires, i.e. when this neuron fires, synapses of every module dependent on it can be updated. Dependencies between the regular and differential modules are satisfied neuron-wise, i.e. according to the actual neurons which have fired in the module, the synapses connected to the respective neurons in the other module can be updated. Synapses of a module dependent on two other modules can only be updated when satisfying both dependency conditions.

[3]The synaptic weight of the reward module is clamped to 1, effectively scaling the other network weights.

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[4] 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). One hundred offsprings were 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 2a). Unsuccessful runs which produce behavior that is not dependent on rewards, show only the first jump (Figure 2b).

In order to assess the conditions for evolving reward-dependent choice behavior (which is indicative of successful reinforcement learning), we also examined a variety of different environmental settings: When bees were evolved

[4]There is no Lamarkian inheritance - learned weights are not passed on to offsprings.

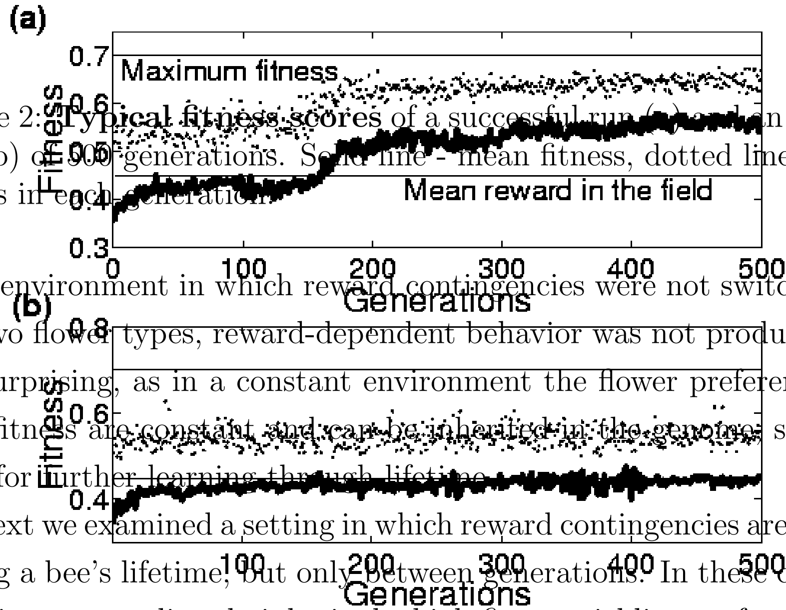


Figure 2: **Typical fitness scores** of a successful run (a) and an unsuccessful run (b) of 500 generations. Solid line - mean fitness, dotted line - maximum fitness in each generation.

in an environment in which reward contingencies were not switched between the two flower types, reward-dependent behavior was not produced[5]. This is not surprising, as in a constant environment the flower preferences yielding high fitness are constant and can be inherited in the genome; so there is no need for further learning through lifetime.

Next we examined a setting in which reward contingencies are not switched during a bee's lifetime, but only between generations. In these conditions an offspring cannot directly inherit the high-fitness-yielding preferences from its parents, as the environment into which it is born may be different from that in which its parents lived. In this scenario we managed to evolve successful reward-dependent behavior. Thus a first condition for the evolution of reward-dependent learning behavior is that reward contingencies must be unpredictable in each generation. This induces the bees to learn to cope with a changing environment. In evolutionary runs in which we not only also switched reward contingencies between generations, but also switched

[5]This was checked by subjecting the evolved bees to tests in environments with different rewarding regimes.

the reward contingencies between the two flower types once during the bees' lifetime, successful reward-dependent behavior evolved much faster. We also found that it is necessary that the timing of the inter-lifetime change in reward contingencies be stochastic (with respect to the trial in which the change occurs), so as not to allow for the evolution of time-related strategies of synaptic weight adjustment, supressing the evolution of reinforcement learning.

We further examined the conditions on the rewarding regimes of the two flower types: In an environment in which both flowers were constant rewarding (but with different amounts of nectar), reward dependent choice behavior was successfully evolved. Thus the *uncertainty between flower types* is a sufficient condition, and *uncertainty within a flower type* (as is the case in most of the simulations hereafter reported, in which one flower is a variably-rewarding flower) is not neccessary for the evolution of reward dependent choice behavior. Moreover, in environments in which both flower types were variably-rewarding types, we were not able to evolve reward-dependent choice behavior. Apparently, in our framework, such excessive uncertainty of the environment is too difficult for the evolutionary process to solve, and the underlying consistencies cannot be discovered and exploited by the evolving bees.

3.2 Exploration/Exploitation Tradeoff

Under the above described conditions in which one flower type is a high-mean constant rewarding flower and the other is a low-mean variably rewarding flower, and in which reward contingencies are switched once during lifetime, about half of the evolutionary runs were successful. Figure 3a 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 the reward, differ-

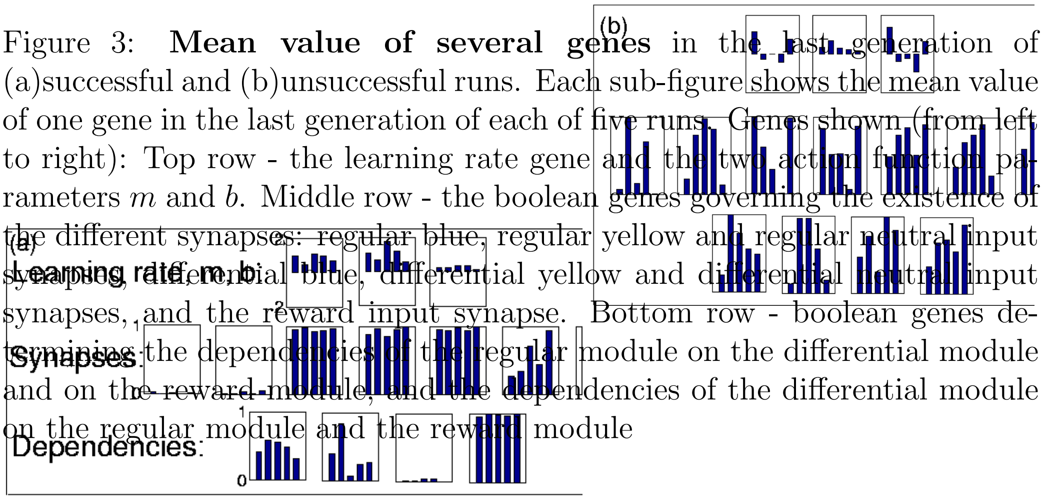


Figure 3: **Mean value of several genes** in the last generation of (a)successful and (b)unsuccessful runs. Each sub-figure shows the mean value of one gene in the last generation of each of five runs. Genes shown (from left to right): Top row - the learning rate gene and the two action function parameters 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

ential 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. Agents which have almost all the crucial alleles, but are missing one or two, are nevertheless unsuccessful (Figure 3b). Thus we find that in our framework, only a network architecture similar to that used by Montague et al. [21] 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.**

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

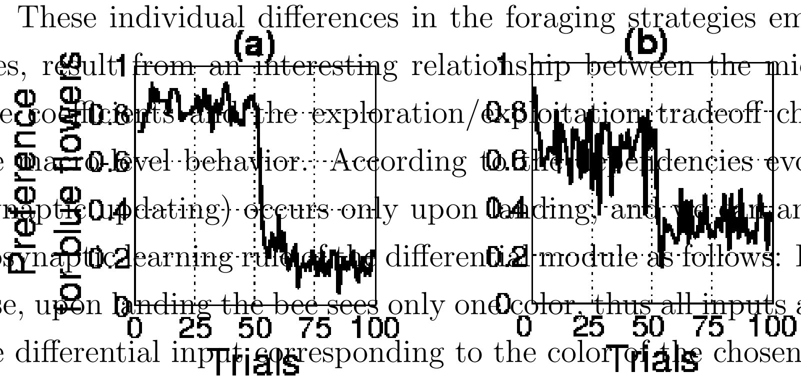


Figure 4: **Preference for blue flowers for two different bees** from the last generation of different successful runs, averaged over 40 test bouts, each consisting of 100 trials. Blue is the initial constant-rewarding high-mean flower. Reward contingencies were switched between flower types at trial 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 $A = -0.92, B = 0.39, C = 0.16, D = 0.25$.

reward contingencies, as they follow the rewards expected from each flower.

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 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[6]. Thus the output of P in this step is:

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

[6]The inputs of the regular module are zero as there is no new input upon landing. Immediately prior to landing the color of the chosen flower filled the bee's field of view, so the differential inputs of the non-chosen colors are zero, and that corresponding to the color of the chosen flower is $X_i(t) - X_i(t-1) = 0 - 1 = -1$.

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

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

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:

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

Thus, a positive D value results in "spontaneous" strengthening of competing synapses (a general rise in the appetitive value of not-visited flower types), leading to an exploration-inclined bee. This behavior is further enhanced by a positive C value, which strengthens competing synapses whenever a "good surprise" (resulting in a positive postsynaptic P value) is encountered. A negative value of D will result in a declining tendency to visit competing flower types as long as the preferred flower does not disappoint (and is thus repeatedly chosen), leading to exploitation-inclined behavior.

3.3 Emergence of Risk Aversion and Matching

A prominent strategy exhibited by the evolved bees is risk-aversion. Figure 6a 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 amounts of nectar, 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 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 [17], we prove analytically that this foraging strategy is a direct consequence of Hebbian learning dynamics in an armed-bandit-like RL situation.

3.3.1 Mathematical Analysis: Risk Aversion is Ordered

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 the bee's stochastic foraging decisions[7] can be formally modeled as choices between a variable-rewarding (v) and a constant-rewarding (c) flower, based on synaptic weights W_v and W_c. For simplicity, let us examine the case of simple monosynaptic anti-Hebbian learning[8]. In this case, the synaptic update rule is the well-known temporal difference (TD) learning rule [28] $\Delta W(t) = \eta(R(t) - W(t-1))$. 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):

$$W_v = W_v(\eta) = \eta(v_t + (1-\eta)v_{t-1} + (1-\eta)^2 v_{t-2} + \cdots) \tag{7}$$

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

In the following we compute f_v, the frequency of visits to variably rewarding flowers. We will 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 as a result, under relatively mild assumptions regarding the bee's

[7]The following analysis relates to the bee's behavior under a certain rewarding regime, i.e. in between changes in reward contingencies.

[8]The monosynaptic part of the evolved Hebbian update rules is in fact anti-Hebbian, as the effective monosynaptic coefficient $(A - C)$ (see equation 5) is approximately (-1) in all successful runs.

choice function, f_v is lower for higher learning rates than for lower learning rates. Thus risk aversion is more prominent with higher 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.**

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[9] [3]) by the expected number of visits to (c) in a typical cycle $[E(n)]$:

$$f_v = \lim_{N\to\infty} \frac{N}{N + \sum_{i=1}^{N} n_i} = \lim_{N\to\infty} \frac{1}{1 + \frac{1}{N}\sum_{i=1}^{N} n_i} = \frac{1}{1 + E(n)} \tag{8}$$

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:

$$E(n) = E[E(n|W_v)] = E\left[\frac{1}{p_v(W_v)} - 1\right] = E\left[\frac{1}{p_v(W_v)}\right] - 1 \tag{9}$$

and so

$$f_v = \frac{1}{E[\frac{1}{p_v(W_v)}]} \tag{10}$$

The mathematical definition of riskiness comes from theories of second degree stochastic dominance [13]. Rothschild and Stiglitz [25] show that for X and Y with a finite equal mean, the following three statements are equivalent:

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

[9] An extension to dependent variables of the Strong Law of Large Numbers.

(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 through for 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.

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:
$$\sum \alpha_i X_i - x = \sum \alpha_i (X_i - x) \leq \sum \alpha_i [\max(X_i - x), 0] \tag{11}$$

Since the right-hand side is non-negative,

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

Now taking expectations of both sides

$$\begin{aligned} E[\max(\sum \alpha_i X_i - x), 0] &\leq \sum \alpha_i E[\max(X_i - x), 0] = \\ &= \sum \alpha_i E[\max(X - x), 0] = \\ &= E[\max(X - x), 0] \end{aligned} \tag{13}$$

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 (7), 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:

$$\begin{aligned} W_v^{(1)}(\alpha) &= \alpha v_1 + \alpha(1-\alpha)v_2 + \alpha(1-\alpha)^2 v_3 + \cdots \\ W_v^{(2)}(\alpha) &= \alpha v_2 + \alpha(1-\alpha)v_3 + \alpha(1-\alpha)^2 v_4 + \cdots \\ &\vdots \\ W_v^{(k)}(\alpha) &= \alpha v_k + \alpha(1-\alpha)v_{k+1} + \alpha(1-\alpha)^2 v_{k+2} + \cdots \end{aligned} \tag{14}$$

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

$$\alpha_1 = \frac{\beta}{\alpha}\ ;\ \alpha_n = \alpha_1(\alpha-\beta)(1-\beta)^{n-2}\ (n \geq 2) \tag{15}$$

We then have:

$$\begin{aligned} \sum_{k=1}^{\infty} \alpha_k W_v^{(k)}(\alpha) &= \frac{\beta}{\alpha}\cdot\alpha\cdot[v_1 + (1-\alpha)v_2 + (1-\alpha)^2 v_3 + \cdots \\ &+ (\alpha-\beta)v_2 + (\alpha-\beta)(1-\alpha)v_3 + (\alpha-\beta)(1-\alpha)^2 v_4 + \cdots \\ &+ (\alpha-\beta)(1-\beta)v_3 + (\alpha-\beta)(1-\beta)(1-\alpha)v_4 + \\ &+ (\alpha-\beta)(1-\beta)(1-\alpha)^2 v_5 + \cdots] = \\ &= \beta[v_1 + (1-\beta)v_2 + (1-\beta)^2 v_3 + \cdots] = W_v(\beta) \end{aligned} \tag{16}$$

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

Now tying this to equation (10) and to statement (i) of the Rothschild and Stiglitz [25] theorem, if $\phi(\cdot) = \frac{1}{p_v(\cdot)}$ is convex (and so $-\frac{1}{p_v(\cdot)}$ is concave), then

$$E\left(-\frac{1}{p_v(W_v(\beta))}\right) \geq E\left(-\frac{1}{p_v(W_v(\alpha))}\right) \tag{17}$$

$$\Rightarrow f_v(\alpha) = \frac{1}{E\left(\frac{1}{p_v(W_v(\alpha))}\right)} \leq \frac{1}{E\left(\frac{1}{p_v(W_v(\beta))}\right)} = f_v(\beta) \tag{18}$$

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

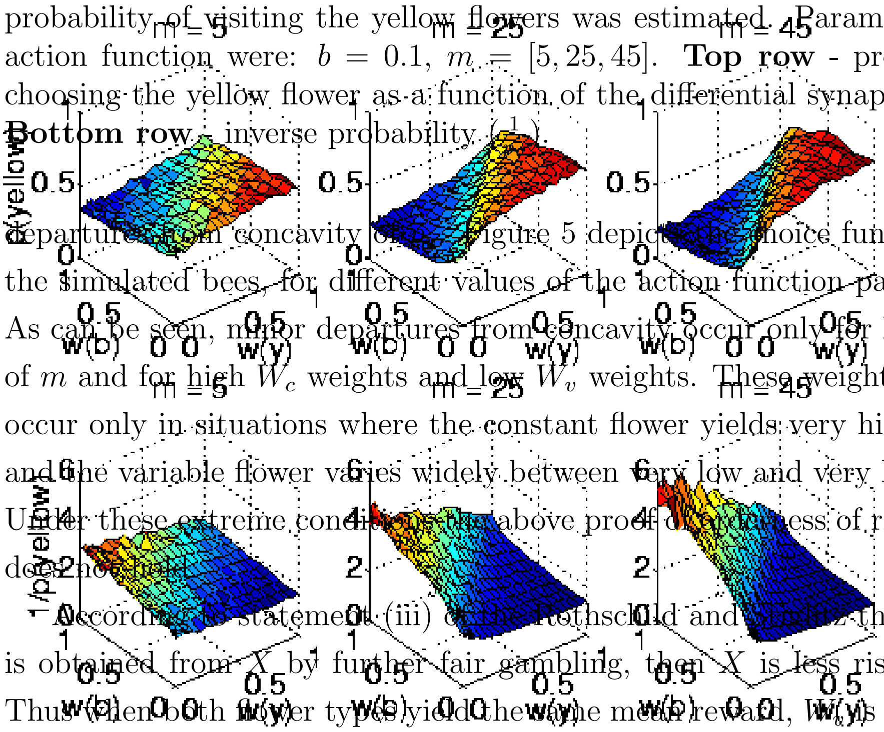


Figure 5: **Choice function of model bee** averaged over 1000 test trials, for different values of m. Blue and yellow differential synaptic weights were clamped to values in the range [0,1], neutral regular weight was clamped to (-1). For each set of weights the bee landed 1000 times from which the probability of visiting the yellow flowers was estimated. Parameters of the action function were: $b = 0.1$, $m = [5, 25, 45]$. **Top row** - probability of choosing the yellow flower as a function of the differential synaptic weights. **Bottom row** - inverse probability [illegible].

departures from concavity [illegible]. Figure 5 depicts the choice function p_v for the simulated bees, for different values of the action function parameter m. As can be seen, minor departures from concavity occur only for large values of m and for high W_c weights and low W_v weights. These weight values can occur only in situations where the constant flower yields very high rewards, and the variable flower varies widely between very low and very high values. Under these extreme conditions the above proof of [illegible]ness of risk-aversion does not hold.

According to statement (iii) of the Rothschild and Stiglitz 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**. Our simulations corroborate these analytical results (Figure 6b).

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 (~ 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 [18], and such abnormally high learning rates have been hypothesized by Real [23], and were also used in Montague et al.'s [21] model. 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.

3.3.2 Probability Matching Behavior

The simulated bees also demonstrate probability-matching behavior. Figure 6(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

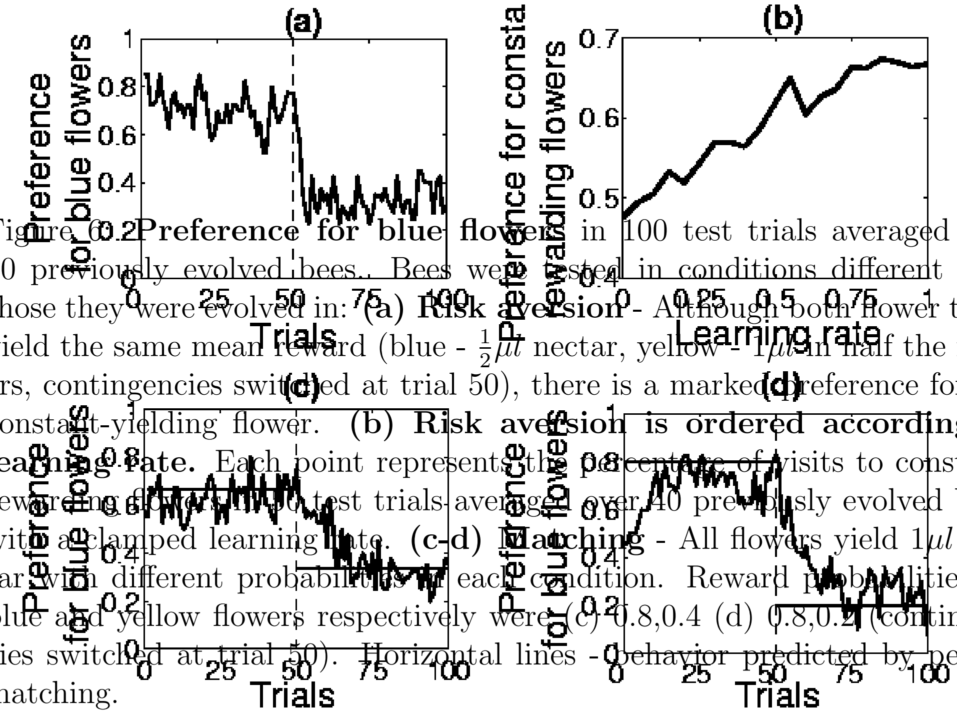


Figure 6.2: **Preference for blue flowers** in 100 test trials averaged over 40 previously evolved bees.. Bees were tested in conditions different from those they were evolved in: **(a) Risk aversion** - Although both flower types yield the same mean reward (blue - $\frac{1}{2}\mu l$ nectar, yellow - $1\mu l$ 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 different probabilities in each condition. Reward probabilities for blue and yellow flowers respectively were (c) 0.8,0.4 (d) 0.8,0.2 (contingencies switched at trial 50). Horizontal lines - behavior predicted by perfect matching.

a direct result of the 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, matching can be evolved in a non-competitive setting, again as a direct consequence of optimal RL.

3.4 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 7). 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.

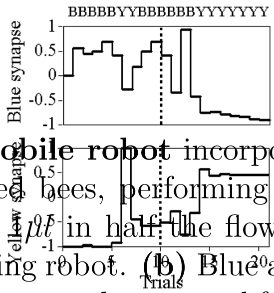


Figure 7: **Synaptic weights of a mobile robot** incorporating a NN controller of one of the previously evolved bees, performing 20 foraging trials (blue flowers - $1/2\mu l$ nectar, yellow - $1\mu l$ in half the flowers, contingencies switched after trial 10). **(a)** The foraging robot. **(b)** Blue 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.

4 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 [1, 14]. 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 [2, 4, 6]. 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 can define necessary and sufficient conditions on the environment, for the successful evolution of reinforcement learning: RL behavior can be evolved in a setting in which both flower types are constant rewarding (with different amounts of nectar), as long as the reward contingencies are not preserved between generations (ie. switched randomly between generations and/or switched in a random timestep during a bee's lifetime). Uncertainty within one of the flower types, although not a necessity for evolving RL behavior, can also be present in the environment, but too much uncertainty in the environment can hamper the ability to produce successful RL: we have not been able to evolve RL behavior in environments in which both flower types were variably-rewarding types, such as in probability matching scenarios[10].

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 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 control the tradeoff between exploration and exploitation.** These behavioral strategies have not been explicitly or implicitly evolved, but emerge in the model as an artifact of optimal learning, 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 the fields of economics and game theory.

In our simulations we find that due to the learning process, the initial weight values are not essential for successful solutions and vary considerably between successful bees. In contrast, the correct learning rule dynamics and the network architecture are crucial for efficient foraging. **These results stress the importance of the encoding of macro level parameters in the genome,** as opposed to encoding neuronal weight values. Such indirect encoding, which is necessary for evolving large networks, can facilitate future research aimed at enlarging the network model by elaborating the visual inputs, as well as the rewarding stimuli. Other future challenges include evolving an action module, which will use the inner reinforcement dopamine-

[10]This is not to say that bees which have been evolved in less uncertain conditions can not subsequentially use their evolved learning mechanism in order to forage successfully in such a scenario, as has been shown here.

like signal produced by P as a basis for the bee's actions in a more complex manner, hopefully producing more complex foraging behaviors.

In summary, the significance of this work is two-fold: on the 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] D. Ackley and M. Littman. Interactions between learning and evolution. In J.D. Farmer C.G. Langton, C. Taylor and S. Rasmussen, editors, *Artificial Life II*. Addison-Wesley, 1991.

[2] J. Baxter. The evolution of learning algorithms for artificial neural networks. In D. Green and T. Bossomaier, editors, *Complex Systems*. IOS Press, 1992.

[3] L. Breiman. *Probability*. Addison-Wesley, 1968.

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

[5] J.W. Donahoe and V. Packard-Dorsel, editors. *Neural network models of cognition: Biobehavioral foundations*. Elsevier Science, 1997.

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

[7] D. Floreano and F. Mondada. Evolutionary neurocontrollers for autonomous mobile robots. *Neural networks*, 11:1461–1478, 1998.

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

[9] U. Greggers and R. Menzel. Memory dynamics and foraging strategies of honeybees. *Behavioral Ecology and Sociobiology*, 32:17–29, 1993.

[10] M. Hammer. An identified neuron mediates the unconditioned stimulus in associative learning in honeybees. *Nature*, 366:59–63, november 1993.

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

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

[13] G.H. Hardy, J.E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, 1934.

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

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

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

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

[18] R. Menzel and U. Muller. Learning and memory in honeybees: From behavior to neural substrates. *Annual reviews in neuroscience*, 19:379–404, 1996.

[19] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[20] P.R. Montague. biological substrates of predictive mechanisms in learning and action choice. In J.W. Donahoe and V. Packard-Dorsel, editors, *Neural network models of cognition: Biobehavioral foundations*, chapter 21, pages 406–421. Elsevier Science, 1997.

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

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

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

[24] L.A. Real. Paradox, performance and the architecture of decision making in animals. *American Zoologist*, 36:518–529, 1996.

[25] M. Rothschild and J. Stiglitz. Increasing risk: I. A definition. *Journal of Economic Theory*, 2:225–243, 1970.

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

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

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

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

[30] T. Unemi, M. Nagayoshi., N. Hirayama, T. Nade, K. Yano, and Y. Masujima. Evolutionary differentiation of learning abilities - a case study on optimizing parameter values in Q-learning by a genetic algorithm. In R.A. Brooks and P. Maes, editors, *Artificial Life IV*, pages 331–336. MIT Press, 1994.