A Neural-Network Model of the Dynamics of Hunger, Learning, and Action Vigor in Mice

Alberto Venditti, Marco Mirolli, Domenico Parisi, Gianluca Baldassarre

Istituto di Scienze e Tecnologie della Cognizione,
Consiglio Nazionale delle Ricerche (ISTC-CNR)
Via San Martino della Battaglia 44, I-00185 Roma, Italy
{*alberto.venditti, marco.mirolli, domenico.parisi, gianluca.baldassarre*}*@istc.cnr.it*

Recently the computational-neuroscience literature on animals' learning has proposed some models for studying organisms' decisions related to the energy to invest in the execution of actions ("vigor"). These models are based on average reinforcement learning algorithms which make it possible to reproduce organisms' behaviours and at the same time to link them to specific brain mechanisms such as phasic and tonic dopamine-based neuromodulation. This paper extends these models by explicitly introducing the dynamics of hunger, driven by energy consumption and food ingestion, and the effects of hunger on perceived reward and, consequently, vigor. The extended model is validated by addressing some experiments carried out with real mice in which reinforcement schedules delivering lower amounts of food can lead to a higher vigor compared to schedules delivering larger amounts of food due to the higher perceived reward caused by higher levels of hunger.

Keywords: Fixed and random ratio schedules, neural networks, average reinforcement learning, motivations, needs, energy costs, phasic and tonic dopamine

1. Introduction

The action of dopamine neuromodulation is believed to exert a powerful influence on *vigor*, that is the strength or rate of responding in behavioural experiments. There are many psychological theories that attribute the vigor effects to a variety of underlying psychological mechanisms, including incentive salience,[1,2] Pavlovian-instrumental interactions,[3,4] and effort-benefit tradeoffs.[5] A different line of research, using the electrophysiological recording of midbrain dopamine neurons's activity in awake behaving monkeys, suggests that the phasic spiking activity of dopamine cells reports to the striatum a specific "prediction error" signal.[6–9] Computational models have shown that this signal can be used efficiently both for learning to predict rewards and for learning to choose actions so as to maximize reward intake.[10–14] However, these theories have some important limitations. First, they only try to explain choice between discrete actions whilst they do not say anything about the strength or vigor of responding. Second, they generally assume that dopamine influences behaviour only indirectly by controlling learning whereas dopamine might have other effects on behaviour. Finally, they are only concerned with the phasic release of dopamine, while the tonic level of dopamine constitutes a potentially distinct channel of neuromodulation that might play a key role in energizing behaviour.[15,16]

Niv et al.[17] proposed a normative account of response vigor which extends conventional reinforcement learning models of action choice to the choice of vigor, that is to the energy expenditure that organisms associate to the execution of chosen actions. To pursue this goal the authors use a model of learning different from the model normally used to study phasic dopamine and reward prediction error, namely the *actor-critic* model based on the *Temporal Difference learning rule*.[18] Rather, they use an actor-critic model based on the *average rate of reward.* The average rate of reward exerts significant influence over overall response propensities by acting as an *opportunity cost* which quantifies the cost of sloth: if the average rate of reward is high, every second in which a reward is not delivered is costly, and therefore actions should be performed faster even if the energy costs of doing so are greater. The converse is true if the average rate of reward is low. In this way the authors show that optimal decision making on vigor leads to choices with the characteristics of choices exhibited by mice and rats in behavioural experiments.

Notwithstanding its pioneering value, the work of Niv et al.[17] has two limits which are addressed here. First, it does not study how food's reinforcing value is influenced by the dynamics of internal needs, e.g. *hunger.* Second, it studies only the steady state values of variables and not their dynamics during learning. This paper proposes a computational model which includes a sophisticated internal regulation of hunger and allows investigating behaviour *during* learning.The results are compared with data from experiments carried out with real mice by Parisi.[19]

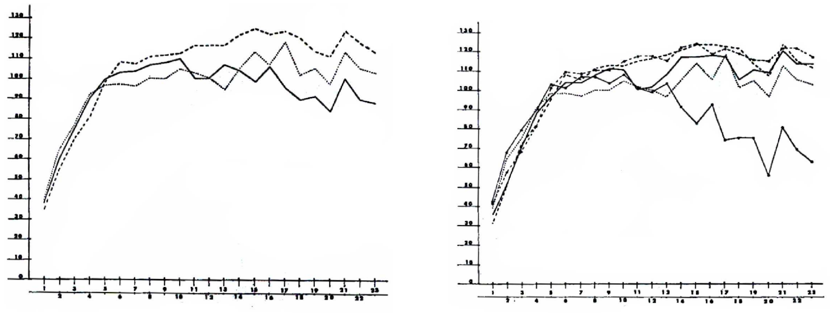
The rest of the paper is organised as follows. Section 2 descrives the targeted experiments. Section 3 illustrates the model and the simulated mice. Section 4 compares the behaviour of simulated and real mice. Finally, Section 5 draws the conclusions.

2. Target experiments

Parisi[19] tested 36 mice in a linear corridor at the end of which they could find a pellet of food, and measured the time they employed to reach the end of the corridor. Here we interpret the speed of mice as an indicator of the vigor invested in the execution of actions. Food was delivered according to three different schedules of reinforcement to three different groups of mice: (a) Fixed Ratio 100% (FR100): food was always delivered when the corridor end was reached. (b) Fixed Ratio 50% (FR50): food was delivered only in odd trials. (c) Random Ratio 50% (RR50): food was delivered randomly with a probability of 50%.

Figure 1a shows the mice's speed curves during learning along various days of training (for each day the average performance for 6 trials is reported). After each daily session the mice had free access to food for half an hour and then were kept without food until the next day session. Figure 1b shows in detail the speed of mice related to FR50 and RR50, separately for trials with and without food (respectively denoted with FR50+ and FR50-. The graphs show that: (a) Mice trained with RR50 exhibited the highest level of vigor, followed by the mice trained with FR100 and then by those trained with FR50 (lowest vigor). The first goal of this paper is to explain why FR100 led to a higher level of vigor with respect to FR50. The high level of vigor exhibited by mice with RR50 was probably caused by some energizing effects of the randomness of action outcomes and will not be further discussed in the paper. (b) Figure 1b shows that FR50+ led to a vigor lower than FR50-. Parisi explained this result suggesting that the reward not only affects learning but it also allows mice to predict the outcome of the succeeding trial (notice that this can happen in FR50, as trials with and without reward alternate and so are predictable, but not in RR50). A second goal of the paper is to validate this hypothesis with the model. (c) Figure 1b also show that FR50+ led to a vigor higher than FR100. At first sight, this is counterintuitive as the reward in FR50+ and FR100 trials is identical. A third goal of the paper, the most important one, is to explain this result in terms of dynamics of *hunger*, namely the fact that higher levels of hunger can increase the perceived reward associated with food. (d) Figure 1b also shows that before vigor levels reach a steady state, FR50- produces the highest levels of vigor, in particular higher than FR50+ and FR100. Parisi explained this by saying that the trials related to FR50- were those taking place right after a rewarded trial (FR+ series). The fourth goal of the paper is to specify and integrate this explanation. Indeed, a further explanation is needed beyond that of Parisi as *both* FR50-

and FR100 conditions involve trials following rewarded trials.



(a)

(b)

Fig. 1. Results of the target experiments. In both graphs, the x-axis refers to successive groups of 6 trials and the y-axis refers to mice's speed (vigor) measured as 100 divided by the number of seconds spent to cover the corridor. (a) The evolution of speed during learning with the three schedules of reinforcement: the highest dashed curve refers to RR50, the intermediate dotted curve to FR100, and the lowest continuous curve to FR50. (b) Same data with separated curves for FR50+ and FR50- (highest dashed curves), and for RR50+ and RR50- (continuous curves); the curve of FR100 is the same.

3. The model

3.1. *The task*

The simulated environment (Figure **??**) is composed by a corridor measuring 1.5 meters. In the experiments, the simulated mouse is placed at the left end of the corridor and is required to decide the speed (vigor) with which to move to the right end. When the mouse reaches the right end it can eventually find (and "eat") a reward (a unit of food) and then is replaced at the start position. The food is delivered according to one of the three reinforcement schedules illustrated in Section 4.

3.2. *The actor-critic component of the model*

The model is based on a neural-network implementation of the actor-critic reinforcement learning model[18] composed of two parts: the *actor* and the *critic* (in its turn mainly formed by the *evaluator*). In general the model is capable of learning to select appropriate actions in order to maximise the *sum of the future discounted rewards*: the evaluator learns to associate evaluations with visited states on the basis of the rewards experienced after these visits; the critic produces a *one-step judgment* of the actor's actions on the basis of the evaluations of couples of states visited in sequence; the actor learns to associate suitable actions with the perceived states of the environment on the basis of the critic's judgment.

This model has been chosen, among the several available reinforcement-learning models, because it has a considerable biological plausibility.[20] In particular, the model has several correspondences with the anatomy and physiology of the basal ganglia, which are deep nuclei of vertebrates' brain playing a fundamental role in action selection.[21]

The model is now illustrated in detail. The model has tree input units: (a) the first two units implement a memory of the outcome, in terms of reward, obtained in the previous trial (in particular, the units are activated with $< 1, 0 >$ when the rat has consumed food in the preceding trial, and with $< 0, 1 >$ otherwise); (b) the third unit is a *bias unit* always activated with 1.

The *actor* is a two-layer feed-forward neural network formed by the three input units, denoted with x_i, and by a sigmoidal output unit ranging in $[0, 1]$ and indirectly encoding vigor. The activation of the output unit is used as the centre μ of a Gaussian probability density function σ having standard deviation ς (set to 0.1) which is used to draw a random number that represents the chosen vigor:

$$\mu = \frac{1}{1 + \exp^{[-\Sigma_i w_{ai} \cdot x_i]}} \qquad y \sim \sigma[\mu, \varsigma] \tag{1}$$

where w_{ai} are the actor's weights from the input units x_i to output unit y and "$\sim$" indicates the probability density function of y (the Gaussian's tails are cut at 0 and 1 by redrawing new numbers when this range is violated). The action y (the selected vigor) is drawn randomly "around μ" as reinforcement learning models need a certain randomness to find suitable solutions by trial-and-error. The activation of the output unit of the actor is used to set the mouse's speed (a maximum vigor of 1 corresponds to a mouse's step size measuring 1/10 of the corridor length).

The *evaluator*, which is part of the *critic*, is a network that uses the activation of the three input units of the model to return, with its linear output unit, an estimation of the theoretical evaluation of the world state corresponding to the input pattern. The "theoretical evaluation" to be estimated, V, is defined as the sum of the future discounted rewards each decreased of the average per-step long-term reinforcement:[22–25]

$$V[t] = E_\pi \left[\sum_{k>t} \left[R[k] - \overline{R} \right] \right] \tag{2}$$

where E_π is the expected sum of future rewards averaged over the possible actions selected by the current action policy π expressed by the current actor, R is the reinforcement, and $\overline{R}$ is the average (per-step) long-term reinforcement. Note that, as suggested by Niv et. al,[17] $\overline{R}$ might be thought to correspond to the tonic dopamine level, encoding the opportunity cost of each time unit engaged in any activity. With this respect, it is important to notice that many experiments show that high levels of striatal dopamine are strongly associated with an high rate of response, that is vigor.[15,16] Interestingly, this happens even before *phasic dopamine* underlying learning (and corresponding to the model's surprise $S[t]$ illustrated below) has a full effect on action selection.[2] In the simulations, $\overline{R}$ is estimated on the basis of the experienced past rewards R. The evaluator produces an estimation $\hat{V}$ of the theoretical V:

$$\overline{R}[t] = (1-\kappa)\overline{R}[t-1] + \kappa R[t] \qquad \hat{V}[t] = \sum_i \left[w_{vi}[t] x_i[t] \right] \tag{3}$$

where w_{vi} are the evaluator's weights ($0 < \kappa < 1$ was set to 0.01).

The *critic* computes the surprise $S[t]$ used to train (as illustrated below) the evaluator to produce increasingly accurate $\hat{V}$ and the actor to produce actions leading to increasingly high and/or frequent rewards:

$$S[t] = \big(R[t] - \overline{R}[t]\big) + \hat{V}[t] - \hat{V}[t-1] \tag{4}$$

The evaluator uses the Temporal Difference algorithm (TD[18]) to learn accurate estimations $\hat{V}$ with experience as follows:

$$w_{vi}[t] = w_{vi}[t-1] + \nu \cdot S[t] \cdot x_i[t-1] \tag{5}$$

where ν is a learning rate (set to 0.2).

The surprise signal is also used by the actor to improve its action policy. In particular, when surprise is positive, the centres of the Gaussian functions used to randomly draw the vigor level are made closer to the actually drawn value, whereas when surprise is negative such centre is moved away" from it. This is done by updating the actor's weights as follows:

$$w_{ai}[t] = w_{ai}[t-1] + \zeta \cdot S[t] \cdot (y[t-1] - \mu[t-1]) \cdot (\mu[t-1](1-\mu[t-1])) \cdot x_i[t-1] \tag{6}$$

where $(\mu[t-1](1-\mu[t-1]))$ is the derivative, with respect to the activation potential, of the actor sigmoid output units' activation, ζ is a learning rate (set to 0.2), $(y[t-1]-\mu[t-1])$ is the part of the formula that moves the centres of the Gaussian towards, or away from, the noisy vigor selected by the actor when surprise $S[t]$ is respectively positive or negative. The motivation behind this updating rule is that a positive surprise indicates that the action randomly selected by the actor at time $t-1$ produced reward effects at time t better than those expected by the evaluator at time $t-1$: this means that such drawn action is better than the "average action" selected by the actor at time $t-1$, as estimated by the evaluator, and so such action should have an increased probability of being selected in the future in correspondence to $x_i[t-1]$. A similar opposite reasoning holds when surprise is negative.

3.3. *The dynamics of costs, hunger, and perceived rewards*

This section illustrates the novel part of the model related to the simulated mouse's energy need (hunger), the energy costs caused by action vigor, the resulting energy balance, and the effects of this on the reward that the mouse perceives when it eats the food. The model of Niv et al.[17] already considered a structure of costs similar to the one illustrated below; however, it did not consider hunger and its effects on perceived rewards, as done here.

In every step, the mouse incurs in two types of costs: (a) a fixed unitary (i.e. per step) cost FUC, set to 0.01; (b) a variable unitary cost VUC set to a maximum level of 0.99: this cost is modulated by the vigor y to capture the fact that more vigor spent executing actions implies a higher energy cost. The sum of the two costs gives the total unitary costs TUC. The energy level E varies on the basis of the energy costs and food ingestion:

$$TUC = FUC + VUC \cdot y^{\iota} \quad E[t] = E[t-1] + \varepsilon \cdot F[t] - \chi \cdot TUC \qquad (7)$$

where ι is a exponential parameter (set to 5.0) implying that costs grow more than proportionally when vigor grows; ε is the energy increases due to the ingestion of one unit of food (set to 0.01), F indicates the units of food ingested when the reward is delivered (set to 10), χ is the decrease of energy due to energy costs (set to 0.05). E is always kept in the range [0, 1]. Moreover, and importantly, at the end of each block of six trials (corresponding to a day session) E is set to 0.2 to represent the fact that after each trial the real mice had free access to food and then were kept without food until the succeeding day session.

Hunger H depends on the level of energy. The perceived reward R, which drives the learning processes of the actor-critic model's components, depends not only on the ingested food but also on the hunger level that modulates the appetitive value of food:

$$H[t] = (1.0 - E[t])^{\varphi} \quad R[t] = F[t] \cdot H[t] \tag{8}$$

where φ is a parameter (set to 3.7) that causes an exponential increase of hunger in correspondence of lower levels of energy.

Figure 2 shows the mouse's costs, perceived rewards, and their balance (difference), all measured per time unit, in correspondence to different levels of vigor and assuming that the mouse starts to run along the corridor with a maximum energy level. The unitary perceived reward UR used to plot the curves was obtained as follows:

$$UR = (F \cdot H) \, / \, (1.5/ \, (MS \cdot y)) \tag{9}$$

where MS is the maximum step size of the mice (set to 1/10 of the corridor length, that is to 0.15), corresponding to the maximum vigor ($y = 1$), and $(1.5/ \, (MS \cdot y))$ represents the number of steps needed by the mice to cover the corridor length (1.5 m) with a vigor y.

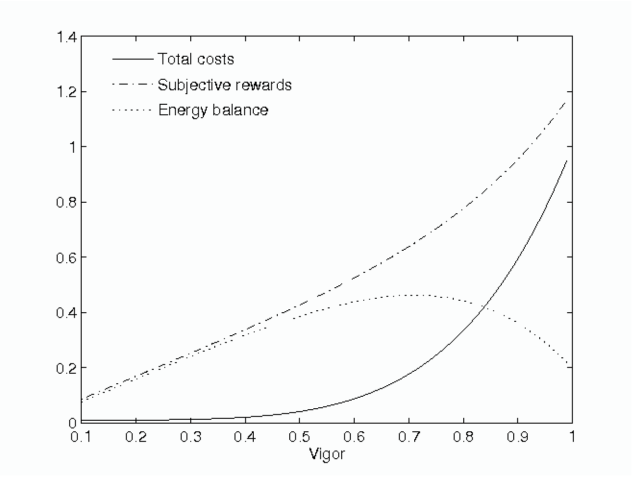


Fig. 2. *The curves represent the energy costs, the perceived rewards, and the energy balance in relation to increasing levels of vigor (x-axis).*

Consider that due to the small duration of a trial the energy spent in terms of TUC is rather low whereas the energy expenditure related to the time that elapses from one day to the following one, which brings E to 0.2 (see above), is rather high and causes the most important effects on perceived rewards. In any case, the dynamics of costs (FUC, VUC and TUC) were included to propose a general model with the potential of tackling many experiments involving hunger. In this respect, it is important to mention that graphs as the one reported in Figure 2, and an analysis of costs as the one reported in this Section, resemble those used by economists to analyse the costs, income and balance of enterprises. Indeed, a recent interdisciplinary research trend in neuroscience aims to exploit the analytical tools used by economics to investigate phenomena related to the functioning of brain.[26] The analysis reported in this section, for example, allowed us to conduct a preliminary exploration of some of the parameters of the model so as to be able to identify interesting regions of them (e.g. this allowed us to envisage the possible vigor value which could maximise the energy balance: see the maximum value of the energy-balance curve in Figure 2).

4. Results

The model was implemented in Java programming language and was tested five times for each of the three reinforcement schedules FR100, FR50 and RR50. The curves of Figure 3 show the level of vigor selected by the model in the three conditions during 10,000 trials of learning. The vigor for FR50 is also plotted for rewarded (FR50+) and non-rewarded (FR50-) trials. These results are now compared with those of Figures 1a-b concerning real mice.

The first result of the simulation is that, as in real mice, with FR100 the simulated mouse selects a level of vigor higher than with FR50. This is likely due to the higher overall energizing effect due to the higher amount of food ingested. More importantly, the model succeeds in reproducing the behaviour of real mice that exhibit a higher vigor with FR50+ than with FR50-: as suggested by Parisi, the reward not only effects learning but, being stored in the model's memory input units, it can also play the role of predictor of the outcome of the next trial.

Figure 1b shows that in real mice FR50+ led to a vigor higher than FR100. As mentioned in Section 2, this result is unexpected as in the two conditions the reward is the same, namely 10 units of food. The model, which reproduces this outcome, allows explaining the mechanism behind it. In the model each group of six trials (corresponding to a "day section" of the experiments with real mice) starts with a level of energy of 0.2. Even if in FR50+ in three trials out of the six of each block the level of energy increases, on average when food is ingested the level of hunger is higher than in FR100. As high levels of hunger increase the *perceived* reward, the mouse learns to spend more energy to get one unit of food in FR50+ than in FR100. Notice how this mechanism might have an adaptive value in ecological conditions as it leads mice to spend more energy when food is scarcer and the risk of death for starvation is high.

Interestingly, the model also reproduces the behaviour exhibited by real mice for which in early phases of learning FR50- produces levels of vigor higher than in the other conditions, in particular FR100 and FR50+. Parisi explained this noticing that the trials related to FR50- were those taking place right after a rewarded trial from FR+. The model suggests detailed mechanisms behind this explanation. According to what stated in the previous paragraph, FR50- trials follow the receiving of the highest perceived reward. In FR50, before the mouse learns to predict if a trial will be rewarded or not the connection weight related to the bias unit will tend to increase maximally in rewarded FR50+ trials and so to contribute to a high vigor in the following FR50- trial. In FR100 this effect is lower as the perceived reward is lower.

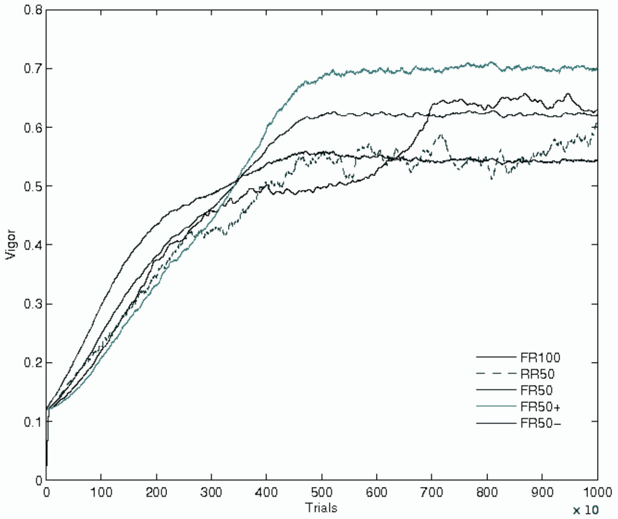


Fig. 3. *Levels of vigor during learning, lasting 10,000 steps, in the conditions FR100, RR50, FR50, FR50+ and Fr50-. Each curve is an average of five repetitions of the simulations.*

5. Conclusion

This paper presented a preliminary study of a model that extends the work of Niv et al.[17] concerning the level of vigor with which animals execute actions by introducing explicitly the dynamics of *hunger* and its effects on *perceived rewards*. The extension makes it possible to reproduce most of the results obtained by Parisi[19] with real mice experiments. The model explains various aspects of the behaviours exhibited by real mice in terms of specific mechanisms, in particular the fact that the vigor of an action can be high even in the presence of low amounts of food received if high levels of hunger lead the mice to perceive food as more rewarding.

The model, however, was unable to reproduce the result according to which real mice trained with RR50 are faster than those trained with FR50 and FR100. As mentioned in Section 2, this particular behaviour is likely due to other mechanisms not taken into consideration by the model, in particular the possible energizing effects of *random* uncertain outcomes. These energizing effects might have adaptive value as they would lead animals to further explore the environment to collect more information and decrease uncertainty. This topic should be addressed in future research.

A second, more important limit of the model, shared with the model of Niv et al.,[17] is that it performs a choice of vigor which is "cognitive", that is, it is learned and implemented on the basis of reinforcement learning mechanisms underlying the selection of actions themselves. On the contrary, probably the nervous system of animals contains some mechanism specifically dedicated to controlling the level of energy invested in actions' performance. One result suggesting that this might be the case is the fact that the model learns to regulate the level of vigor very slowly (in about 4,000 trials) while real mice regulate the level of vigor after few trials, often even before they learn to produce the correct action.[17] Also this issue should be tackled in future work.

Acknowledgments

This research was supported by the EU Project *ICEA - Integrating Cognition, Emotion and Autonomy*, contract no. FP6-IST-IP-027819. Venditti, Mirolli, and Baldassarre consider this research also a small homage to the bright and curious mind of the young Domenico Parisi which, 43 years ago, in times of behaviourism dominance, led him to go abroad in USA to carry

out the experiments addressed in the paper. This experience was one of the first forward-running steps of a long, always-innovative, enthusiastic scientific career which ultimately arrived to mark forever the life of the co-authors and other dozens of young scientists.

References

1. K. C. Berridge and T. E. Robinson, *Brain Res Brain Res Rev* **28**, 309 (1998).
2. S. Ikemoto and J. Panksepp, *Brain Res Brain Res Rev* **31**, 6 (1999).
3. A. Dickinson, J. Smith and J. Mirenowicz, *Behav Neurosci* **114**, 468 (2000).
4. A. Murschall and W. Hauber, *Learn Mem* **13**, 123 (2006).
5. J. D. Salamone and M. Correa, *Behav Brain Res* **137**, 3 (2002).
6. T. Ljungberg, P. Apicella and W. Schultz, *J Neurophysiol* **67**, 145 (1992).
7. W. Schultz, P. Apicella and T. Ljungberg, *J Neurosci* **13**, 900 (1993).
8. W. Schultz, *J Neurophysiol* **80**, 1 (1998).
9. P. Waelti, A. Dickinson and W. Schultz, *Nature* **412**, 43 (2001).
10. R. S. Sutton and A. G. Barto, *Psychol Rev* **88**, 135 (1981).
11. K. J. Friston, G. Tononi, G. N. Reeke, O. Sporns and G. M. Edelman, *Neuroscience* **59**, 229 (1994).
12. A. G. Barto, *Curr Opin Neurobiol* **4**, 888 (1994).
13. P. R. Montague, P. Dayan and T. J. Sejnowski, *J Neurosci* **16**, 1936 (1996).
14. W. Schultz, P. Dayan and P. R. Montague, *Science* **275**, 1593 (1997).
15. G. D. Carr and N. M. White, *Pharmacol Biochem Behav* **27**, 113 (1987).
16. D. M. Jackson, N. E. Anden and A. Dahlstroem, *J Psychopharmacol* **45**, 139 (1975).
17. Y. Niv, N. D. Daw, D. Joel and P. Dayan, *J Psychopharmacol* **191**, 507 (2007).
18. R. Sutton and A. Barto, *Reinforcement Learning: An Introduction.* (MIT Press, Cambrige, MA, USA, 1998).
19. D. Parisi, Il rinforzo come stimiolo discriminativo, in *Atti del XV Congresso degli Psicologi Italiani*, 1965.
20. J. Houk, J. Davis and D. Beiser, *Models of Information Processing in the Basal Ganglia* (MIT Press, Cambridge, MA, USA, 1995).
21. P. Redgrave, T. J. Prescott and K. Gurney, *Neuroscience* **89**, 1009 (1999).
22. A. Schwartz, A reinforcement learning method for maximizing undiscounted rewards, in *Proceeding of the Tenth Annual Conference on Machine Learning*, 1993.
23. S. Mahadevan, *Machine Learning* **22**, 159 (1996).
24. J. Tsitsiklis and B. V. Roy, *Automatica* **35**, 1799 (1999).
25. N. D. Daw and D. S. Touretzky, *Neural Comput* **14**, 2567 (2002).
26. P. Phillips, M. Walton and T. Jhou, *J Psychopharmacol* **191**, 483 (2007).