

A Computational Model of Hunger, Perceived Reward and Vigor in Experiments of Operant Conditioning with 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.paris, gianluca.baldassarre}@istc.cnr.it
www.laral.istc.cnr.it*

Recently the computational-neuroscience literature on animals' learning has proposed some models to study organisms' decisions related to the energy to invest in the execution of actions ("vigor"). These models are based on average reinforcement learning algorithms which allow reproducing organisms behaviours and at the same time allow linking 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 expenditure and food ingestion, and the effects of it on perceived reward and consequently vigor. The extended model is validated by addressing some experiments carried out with real mice where reinforcement conditions using lower amounts of reward can lead to a higher vigor with respect to conditions using larger amounts of reward due to the increase of the perceived appetitive value of reward.

Keywords: Fixed random ratio reinforcement schedules, neural networks, average reinforcement learning, motivations, needs, energy costs, phasic 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 works, related to the electrophysiological recordings 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 showed 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} These computation theories have some important limits: firstly, they only take into account the choice between discrete actions whilst they do not say anything about the strength or vigor of responding. Secondly, they generally assume that dopamine influences behaviour only indirectly by controlling learning. 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 behavior.¹⁵

In these background of works, Niv et al.¹⁵ proposed a normative account of response vigor which extends the 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 that usually used to study phasic dopamine and reward prediction error, namely the *actor-critic* model based on the *Temporal Difference learning rule*.¹⁶ 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 psychological experiments.

Notwithstanding its pioneering value, the work of Niv et al.¹⁵ has two limits which are addressed here. First, their model does not study how the reinforcing value of food is influenced by the dynamics of internal needs, e.g. *hunger*. Second, the work studies only the steady state values of variables and not the dynamics of the learning processes. This paper proposes a computational model that starts to overcome these limits in that it studies a system with a sophisticated internal regulation of hunger, and starts to investigate behaviour *during* learning on the basis of data from some real experiments carried out with real mice by Parisi.¹⁷

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

2. Target experiments

Parisi¹⁷ tested 36 mice in a linear corridor and measured the time that they employed to cover it from one end to the second end where they could eventually find some food (here we interpret the speed of mice as an indicator of the vigor invested in the execution of actions). The food was delivered with three different schedules of reinforcement to three sub-groups of mice. These schedules were: (a) Fixed Ratio 100% (FR100): the food was always delivered when the mice reached the second end of the corridor. (b) Fixed Ratio 50% (FR50): when the mice reached the second end of the corridor the food was delivered only in odd trials. (c) Random Ratio 50% (RR50): when the mice reached the second end of the corridor the food was delivered randomly with a probability of 50%.

Figure 1 shows the results of the experiments. Figure 1a reports the mice's vigor (speed) curves during learning along various days of training (each day reports the average performance for 6 trials; after each day session the mice had free access to food for half an hour and then they were kept without food until the succeeding day session). Figure 1b shows in detail the speed of mice related to FR50 and RR50 referring to trials with and without food at the end of the corridor (respectively denoted with FR50+ and FR50-): the resulting curves allow analysing the effects on vigor of trials ending with and without reward. Various remarkable facts are apparent from these graphs: (a) Mice trained with with RR50 exhibited the highest level of vigor, followed by the mice trained with FR100, which exhibited an intermediate vigor, and then by those trained with FR50, which exhibited the lowest level of vigor. One of the 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, probably related to some energizing effect of the randomness of action outcomes, will not be not tackled here. (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 allow 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, where mice do not have any information on the possible outcome of the trials, see Figure 1b). A second goal of this paper is to validate this hypothesis with the proposed computational model. (c) Figure 1b also show that FR50+ led to a vigor higher than FR100. At first sight, this is a counterintuitive result as the reward in FR50+ and FR100 trials is identical. A third goal of this paper, the most important, is explaining this result in terms of dynamics of *hunger*, namely the fact that

higher levels of hunger can increase the perceived rewarding value of food. (d) Figure 1b also shows that before vigor levels reach a steady state, FR50- produces the highest levels of vigor. 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 need beyond that of Parisi as *both* these conditions involve trials following rewarded trials.

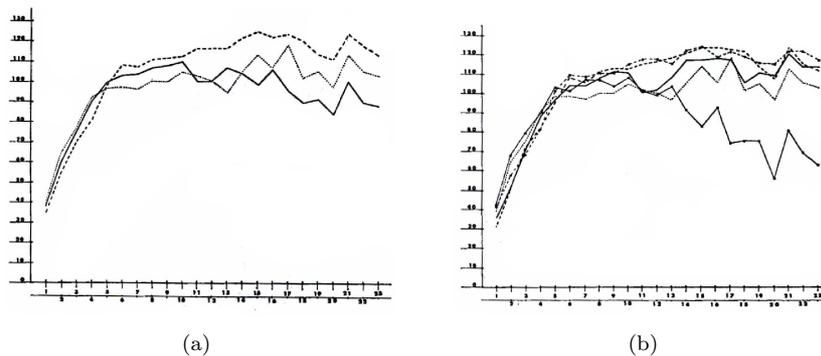


Fig. 1. Results of the target experiments. In both graphs, the x-axis reports groups of 6 trials whereas the y-axis reports the mice's speed (vigor) measured as 100 divided by the time spent to cover the corridor measured in seconds. (a) The evolution of speed during learning with the three schedules of reinforcement: the curve with the highest steady state (fastest mice, best performance: dashed line) refers to the condition RR50, the curve with the intermediate steady state (intermediate performance, dotted line) refers to FR100, and the curve with the lowest steady state (lowest performance, continuous line) refers to FR50. (b) Details of FR50 and RR50 obtained by measuring separately the performance of mice when they found the food at the end of the corridor (FR50+) and when they did not found it (FR50-).

3. The model

3.1. The task

The simulated environment (Figure 2) 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 corridor where it can eventually obtain a reward. When the mouse reaches the right end of the corridor it can find a reward (a unit of food). The food is delivered according to one of the three reinforcement schedules employed in the experiments with real mice and

illustrated in Section 4. The simulation takes place in discrete time steps: a time step involves an input/output cycle of the neural network controlling the mouse and the execution of the mice's actions. When the mouse enters the rewarded end of the corridor, it "eats the food", if the food is there, and then is replaced at the start position.

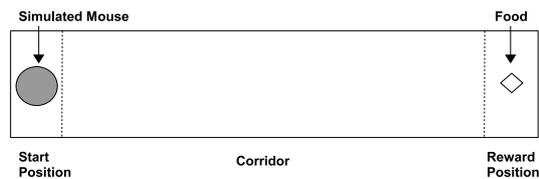


Fig. 2. *The simulated experimental setup.*

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

The model is based on a neural network whose architecture is presented in Figure 3. The system is a neural-network implementation of the actor-critic reinforcement learning model.¹⁶ The model is mainly composed of two parts: the *actor* and the *critic* (on 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 single 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.¹⁸ In particular, the model has several correspondences with the anatomy and physiology of basal ganglia, some deep nuclei of vertebrates' brain playing a fundamental role in action selection.¹⁹ In this respect, computations similar to those performed by the actor might be implemented by the portion of the *striatum* (the input component of the basal ganglia) named *matrix*, involved in the selection of actions. Moreover, the computations similar to those performed by the evaluator might be implemented by the portions of the striatum named *striosomes*, that play an important role, via dopamine cells

(located in the ventral tegmental area and substantia nigra pars compacta), in the learning processes of basal ganglia.

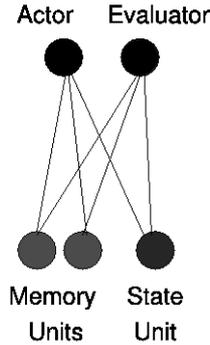


Fig. 3. The architecture of the neural controller of the simulated mouse.

These model's functioning is now illustrated in details. The model has three input units. The first two input units implement a memory of the outcome, in terms of reward, obtained in the last trial. In particular, these units are activated with $\langle 1, 0 \rangle$ or $\langle 0, 1 \rangle$ in the case the rat has consumed or not food in the last trial. The third input 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 the vigor. In this respect, 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) used to draw a random number that represents the chosen vigor:

$$\mu = \frac{1}{1 + \exp[-\sum_i w_{ai} \cdot x_i]} \quad (1)$$

$$y \sim \sigma[\mu, \varsigma] \quad (2)$$

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 , corresponding to the selected vigor, is drawn randomly

“around μ ” as reinforcement learning models need a certain randomness to explore new actions and find suitable solutions by trial and error. The activation of the output unit of the actor is used to set the mice’s speed. To this purpose, a maximum vigor of 1 is assumed to correspond to a mouse’s step size measuring 1/10 of the corridor length.

The *evaluator*, 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:²⁰⁻²³

$$V[t] = E_{\pi} \left[\sum_{k>t} [R[k] - \bar{R}] \right] \quad (3)$$

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 \bar{R} is the average (per-step) long-term reinforcement. Note that, as suggested by Niv et. al,¹⁵ \bar{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.^{24,25} 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, \bar{R} is estimated on the basis of the experienced past rewards R :

$$\bar{R}[t] = (1 - \kappa)\bar{R}[t - 1] + \kappa R[t] \quad (4)$$

where $0 < \kappa < 1$ (κ was set to 0.01).

The evaluator produces an estimation \hat{V} of the theoretical evaluation V :

$$\hat{V}[t] = \sum_i [w_{vi}[t]x_i[t]] \quad (5)$$

where w_{vi} are the evaluator’s weights.

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:

$$S[t] = (R[t] - \bar{R}[t]) + \hat{V}[t] - \hat{V}[t - 1] \quad (6)$$

The evaluator uses the Temporal Difference algorithm (TD¹⁶) to learn accurate estimations \hat{V} with experience as follows:

$$w_{vi}[t] = w_{vi}[t - 1] + \nu \cdot S[t] \cdot x_i[t - 1] \quad (7)$$

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:

$$w_{ai}[t] = w_{ai}[t-1] + \zeta \cdot S[t] \cdot (y[t-1] - \mu[t-1]) \cdot (\mu[t-1](1 - \mu[t-1])) \cdot x_i[t-1] \quad (8)$$

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 way of updating the actor’s weights 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 same condition. A similar, but opposite, reasoning holds when surprise is negative.

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

This section illustrates the novel part of the model related to the simulated mouse’s energy need (hunger), the energy costs caused by the vigor spent in executing the actions, the resulting energy balance, and the effects of

this on the reward perceived by eating the food. In particular, the model of Niv et al.¹⁵ 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.

Let us start to consider the structure of the energy costs of the simulated mouse. In every step, the simulated mouse incurs in two types of costs: (a) a fixed unitary 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 expenditure. The sum of the two sources of costs given the total unitary costs TUC per step:

$$TUC = FUC + VUC \cdot y^\iota; \quad (9)$$

where ι is a exponential parameter (set to 5.0) implying that costs grow more than proportionally when vigor grows.

The energy level E varies on the basis of the energy costs and food ingestion:

$$E[t] = E[t-1] + \varepsilon \cdot F[t] - \chi \cdot TUC \quad (10)$$

where 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 each block of six trials (corresponding to a day session of the original experiment) 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 experimental day session.

Hunger H depends on the level of energy as follows:

$$H[t] = (1.0 - E[t])^\varphi; \quad (11)$$

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

$$R[t] = F[t] \cdot H[t]; \quad (12)$$

Figure 4 shows the mice's costs, perceived rewards, and their balance (difference), all measured per time unit, in correspondence to different levels of vigor and assuming that the mice start to run along the corridor with a maximum energy level. The plot the curves, the unitary perceived reward UR has been obtained as follows:

$$UR = (F \cdot H) / (1.5 / (MS \cdot y)); \quad (13)$$

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 .

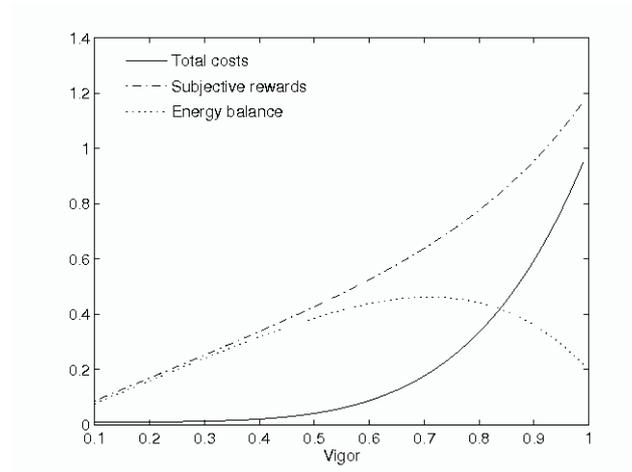


Fig. 4. The curves represent the energy costs, the perceived rewards, and the energy balance in relation to increasing levels of vigor (x -axis).

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 4, 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 4).

4. Results

The model was implemented in Java programming language and was tested with the three schedules of reinforcement FR100, FR50 and RR50 (the simulations were repeated five times per condition).

The curves of the Figure 5 show the level of vigor in the mice of the three conditions during 10,000 trials of learning. The performance related to condition FR50 is also plotted by dividing the trials in those rewarded (FR50+) and those not rewarded (FR50-). These results are now compared with those illustrated in Figures 1a-b related to the outcome of the experiments run with real mice.

The first result of the simulation is that, as in real mice, FR100 causes a level of vigor higher than 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-. The explanation of this is that, as suggested by Parisi, the reward not only effects learning but it can also play the role of predictor of the outcome of the succeeding trial. In the model, this fact is captured by the memory input units that allow the simulated mice to regulate vigor on the basis of the outcome of the previous trial.

Figure 1b shows that FR50+ led to a vigor higher than FR100 in real mice. As mentioned in Section 2, at first sight 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 fact, 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 the three trials out of the six of each block the level of energy increases, on average the level of hunger when food is ingested is higher than when food is ingested in the FR100 condition. As high levels of hunger increase the *perceived* reward, the result is that the

mice will learn to spend more energy to get one unit of food in FR50+ with respect to FR100. Notice how this mechanism might have an adaptive value in ecological conditions as it leads the mice to spend more energy when food is scarce and the risk of starvation death increases, than when food is abundant.

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 saying that the trials related to FR50- were those taking place right after a rewarded trial (FR+ series). The model suggests detailed mechanisms behind this explanation. According to what stated in the previous paragraph, FR50- trials follow the receipt of the highest perceived reward. In FR50, before the mice learn to predict if a trial is rewarded or not on the basis of the memory of the previously obtained reward, 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- trials. In FR100 this effect is lower as the perceived reward is lower.

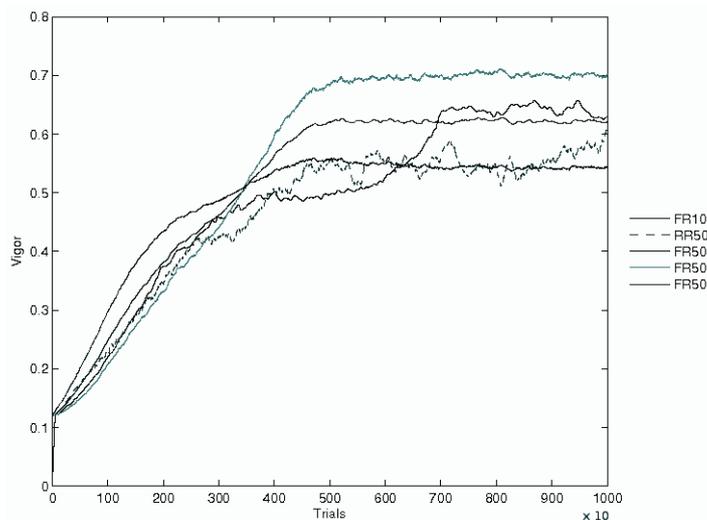


Fig. 5. Levels of vigor during learning, lasting 10,000 steps, in the conditions FR100, RR50, FR50, FR50+ and Fr50- (average of five runs for each condition). Each curve is an average of five repetitions of the simulations.

Last, the simulations also show that the model was unable to reproduce the result obtained with real mice for which the group trained with RR50 is

the fastest among the three groups. As mentioned in Section 2, in real mice this particular behaviour is likely due to aspects not taken into consideration by the model, for example the possible energizing effects of *random* outcomes of trials which might trigger a higher activity in order to explore the environment to collect further information and decrease uncertainty. This topic might be addressed in future research.

5. Conclusion

This paper presented a preliminary study of a model that extends the work of Niv et al.,¹⁵ related to the decision of the level of vigor of performed actions, by introducing the dynamics of *hunger* and its effects on *perceived rewards*.

This extension allowed reproducing some results obtained in the experiments conducted by Parisi¹⁷ with real mice. Moreover, the model allowed to explain in terms of specific mechanisms various aspects of the behaviors exhibited by real mice in the targeted experiments, in particular the fact that the vigor of action can be high even in the presence of low amounts of food received if high levels of hunger lead the mice to perceive them as more rewarding.

One important limit of the model, shared with the model of Niv et al.,¹⁵ is that the choice of vigor is somehow “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 a “machinery” specifically dedicated to control the level of energy invested in actions’ performance. One empirical data suggesting this is the fact that the model presented here learns to regulate the levels of vigor very slowly (in particular in about 4,000 trials) while real mice regulate the levels of vigor after few trials, often even before learning to produce the correct action.¹⁵ This issue might 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.

References

1. K. C. Berridge and T. E. Robinson, *Brain Res Brain Res Rev* **28**, 309(Dec 1998).
2. S. Ikemoto and J. Panksepp, *Brain Res Brain Res Rev* **31**, 6(Dec 1999).

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