Behavioural Brain Research 198 (2009) 125–129

Contents lists available at ScienceDirect

Behavioural Brain Research

journal homepage: www.elsevier.com/locate/bbr

Research report

Theory meets pigeons: The influence of reward-magnitude on discrimination-learning

Jonas Rose[a,*], Robert Schmidt[b,c], Marco Grabemann[a], Onur Güntürkün[a]

[a] Institute of Cognitive Neuroscience, Biopsychology, Department of Psychology, Ruhr-University of Bochum, 44780 Bochum, Germany
[b] Institute for Theoretical Biology, Department of Biology, Humboldt-Universität zu Berlin, Invalidenstr. 43, 10115 Berlin, Germany
[c] Bernstein Center for Computational Neuroscience Berlin, Philippstr. 13, 10115 Berlin, Germany

ARTICLE INFO

Article history:
Received 22 August 2008
Received in revised form 14 October 2008
Accepted 20 October 2008
Available online 8 November 2008

Keywords:
Temporal-difference
Reinforcer magnitude
Learning-rate
Animal behavior
Reinforcement learning

ABSTRACT

Modern theoretical accounts on reward-based learning are commonly based on reinforcement learning algorithms. Most noted in this context is the temporal-difference (TD) algorithm in which the difference between predicted and obtained reward, the prediction-error, serves as a learning signal. Consequently, larger rewards cause bigger prediction-errors and lead to faster learning than smaller rewards. Therefore, if animals employ a neural implementation of TD learning, reward-magnitude should affect learning in animals accordingly.

Here we test this prediction by training pigeons on a simple color-discrimination task with two pairs of colors. In each pair, correct discrimination is rewarded; in pair one with a large-reward, in pair two with a small-reward. Pigeons acquired the 'large-reward' discrimination faster than the 'small-reward' discrimination. Animal behavior and an implementation of the TD-algorithm yielded comparable results with respect to the difference between learning curves in the large-reward and in the small-reward conditions. We conclude that the influence of reward-magnitude on the acquisition of a simple discrimination paradigm is accurately reflected by a TD implementation of reinforcement learning.

1. Introduction

Successful behavior depends on establishing reliable predictions about future events. To select appropriate actions, humans and other animals need to learn which sensory events predict dangers or benefits and which actions improve or worsen the situation of the animal. This learning often relies on positive (reward) or negative feedback (punishment). The neural basis of feedback-based learning is highly conserved across species and much of the basic neural organization in different vertebrate species resembles each other [38,12]. Countless research has been dedicated to understanding the computational principles mediating feedback-based learning and numerous models have been devised to describe these principles mathematically [36,8]. Modern, theoretical accounts on feedback-based learning are mostly centered on reinforcement learning algorithms; the most prominent of these is the temporal-difference (TD) algorithm [36,37], which has been successfully used as a model for behavioral and neural responses during reward-based learning [21,31]. TD learning is an extension of the Rescorla–Wagner (or also the Widrow–Hoff) learning rule, with a more detailed representation of time [36,37]. We used the TD model in this study because it is widely used in computational neuroscience and because it is well integrated into machine-learning theory including action selection in decision making.

In TD-algorithms, time is often divided into discrete steps and for each time step the amount of predicted future reward is determined on the basis of sensory stimuli. A comparison of predicted and obtained reward yields a prediction error signal with three basic characteristics: (1) an unexpected reward generates a positive prediction error indicating that more reward was obtained than was predicted, (2) omission of a predicted reward generates a negative prediction error indicating that less reward was obtained than was predicted, and (3) obtaining a fully predicted reward generates no prediction error. This prediction error signal is in turn used to update the reward prediction of sensory stimuli that preceded the reward; a positive prediction error leads to an increase in reward prediction, a negative prediction error to a decrease in reward prediction [31,33]. Through these mechanisms TD learning can be used to associate a stimulus with a reward (as in classical conditioning) [25], to associate an action with a reward (as in operant conditioning) [22,1] or also to cause extinction of a previously formed association [26].

The TD-algorithm gained popularity, since the activity of dopaminergic neurons located in the ventral tegmentum and substantia nigra pars compacta of mammals resembles the TD prediction error signal. The dopaminergic system is frequently termed

* Corresponding author. Tel.: +49 2343226845; fax: +49 23432214377.
E-mail address: jonas.rose@rub.de (J. Rose).

0166-4328/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbr.2008.10.038

the 'reward-system' of the brain and numerous theories have been devised on its exact role in reward. The most prominent theories include reinforcement [35], incentive salience [2] and habit formation [10]. Despite the discussion on the behavioral role of dopamine, there is clear evidence that the activity of dopaminergic neurons bears striking resemblance to the TD error signal. The responses of dopaminergic neurons show positive and negative prediction errors [21,31,25] and comply with several assumptions of learning theory [40]. One important prediction of the TD-algorithm is that the error signal is dependent on the size of the reward; a big unexpected reward will generate a bigger error signal than a small unexpected reward. Hence, bigger rewards lead to faster learning than smaller rewards.

The influence of reward-magnitude on animal behavior has previously been investigated with regards to several questions, for example reward-discriminability [7,14,15,17,24], motivation [4–6,9,19,43] and choice behavior [18]. In addition, it has been evaluated in the light of response-rates during acquisition [4,7,13,20,43], and reversal [19]. However, whether the influence of reward-magnitude on learning-rate complies with the predictions of the TD-model has not yet directly been investigated. Such a test requires the use of error-rates instead of measures of response-strength in order to avoid measuring overall differences in performance due to motivational differences [5,6]. Here we test whether the acquisition of a color-discrimination is modulated by the magnitude of contingent reward and relate our findings to an implementation of the TD-model.

2. Materials and methods

2.1. Subjects

Twelve naive homing pigeons (*Columba livia*) with body weights ranging from 330 g to 490 g served as subjects. The animals were housed individually in wire-mesh cages inside a colony room, had free access to water and grit and during experiments they were maintained on 80% of their free-feeding body weight. The colony room provided a 12 h dark–light cycle with lights on at 8:00 and lights off at 20:00. The experiment and all experimental procedures were in accordance with the National Institute of Health guidelines for the care and use of laboratory animals and were approved by a national committee (North Rhine-Westphalia, Germany).

2.2. Apparatus and stimuli

All training and testing was conducted in an operant chamber, controlled via PC and parallel-port interface by Matlab (the Mathworks Inc.) and the Biopsychology Toolbox [29]. Situated on the front panel of the chamber were four pecking keys, transparent, circular switches of 2.5 cm diameter, behind these was a TFT-Monitor (Acer AL1511) used for presentation of the stimuli. Two pecking keys were placed on the sides, 14 cm above the two feeders; the other keys were placed centrally, one above the other (8 cm distance, the lower key 18 cm above the floor). The stimuli consisted in a full back-illumination of a given pecking key, either in white or in one of four basic colors (red, green, blue, yellow). These stimuli were always presented in the combinations red–green and blue–yellow, one color of each pair serving as S+, the other as S−. For each bird, one combination was paired with the chance of gaining a large-reward, the other with the chance of gaining a small-reward. For each animal one feeder gave access to grain for 4.0 s and the other for 1.5 s these served as large- and small-rewards, respectively. Mixed grain was used as reward. All contingencies (the color of the S+, color-pair and reward-size, reward-size and side of the reward) were balanced between the animals.

2.3. Behavioral task

The birds were trained on two distinct tasks, on a simple discrimination between a large- and a small-reward and on a simple discrimination of basic colors. During pre-training the animals were trained in an autoshaping procedure to respond to the pecking keys, thereafter they were trained on an operant conditioning (FR1) schedule. The series of events was similar in both paradigms, after an inter-trial interval of 10 s the left or right pecking key was illuminated in white for 9 s. A peck to the illuminated key resulted in a reward delivered by the feeder situated below the pecking key. For each animal, one feeder always delivered the large-reward, giving access to food for 4.0 s, the other always delivered the small-reward, giving access to food for 1.5 s; the side of the 'good' and the 'bad' feeder was balanced between animals. Omission of a response was rewarded in the autoshaping-trials but mildly punished with 10 s lights off in the FR1-trials. When the animals showed stable pecking-responses, training on the reward-choice commenced (Fig. 1). In these trials, an inter-trial interval of 10 s was followed by the illumination of both side-keys in white light. Response to either key was rewarded by the corresponding feeder below it. Choice of the side-key thus determined the feeder that delivered a reward and consequently the duration of access to food. Omission of a response was mildly punished with 10 s lights off.

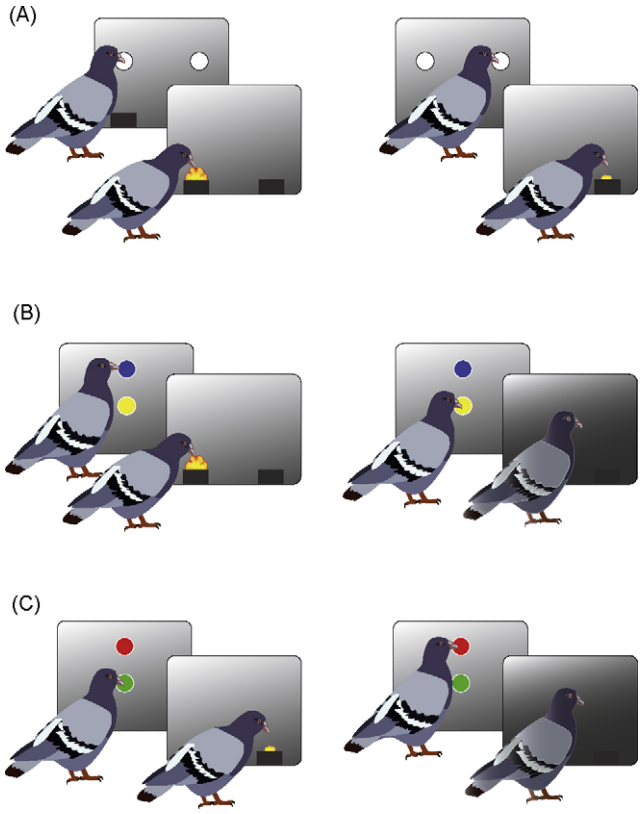

Fig. 1. The behavioral task. (A) Reward-choice trial, the animals learn to associate a side-key with the corresponding feeder and the corresponding reward-magnitude. A response on the left key will result in reward-delivery on the left feeder, each animal has a 'good' and a 'bad' feeder that will always deliver the large- and small-rewards, respectively. (B) A color-choice trial with large-reward; choice of the S++ (blue key) results in the large-reward, choice of the S− (yellow key) and response omission result in a mild punishment. (C) A color-choice trial with small-reward; choice of the S+ (green key) results in a small-reward, choice of the S− (red key) and response omission result in a mild punishment. All contingencies are balanced between the animals.

After the criterion, three consecutive days with at least 80 percent choice of the large-reward, was reached the animals were trained on the color-discrimination (Fig. 1). In each session, reward-discrimination trials and color-discrimination trials were presented in a block-wise fashion (10 trials reward-, 20 trials color-, 10 trials reward-, 20 trials color-discrimination). Color-discrimination trials were separated by a 10 s inter-trial interval after which the choice stimuli were presented on the central pecking keys. Correct response to S++ resulted in a large-reward, correct response to the S+ in a small-reward. Response to S− and omission of a response were mildly punished with 10 s lights off.

2.4. Data analysis

Animal behavior was analyzed with respect to differences in learning on the large-reward and small-reward conditions. All analysis was performed using the percentage of correct trials in a session. Two distinct measures were used: a direct comparison of the learning curves and the number of sessions required to reach criterion. The direct comparison was performed using a Wilcoxon signed-rank test. The other measure, sessions to criterion, was evaluated for a criterion of 75% correct. A paired Student's *t*-test was used for significance-testing between the conditions.

The comparison of behavioral and modeling data was performed with respect to the difference in learning on the large-reward and learning on the small-reward conditions. For this comparison, the mean performance of all animals for small-reward and large-reward stimuli was calculated for each day. The small-reward curve was then subtracted from the corresponding large-reward curve. Performing this calculation resulted in a difference-curve for the behavioral data and one for the

modeling data. Correlation coefficients were determined between these curves. This approach allowed comparing the influence of reward-magnitude on learning-rates in the behavioral date with the influence reward-size had in the model.

2.5. Modeling

To see whether the behavioral data matches the predictions of canonical models of animal behavior we implemented a reinforcement learning algorithm. We used a standard actor-critic architecture that employs TD learning.

The critic component learns to predict future rewards on the basis of sensory stimuli. A complete serial compound stimulus representation was chosen in which the occurrence of a stimulus was represented in a state vector s. Usually, this vector contains only zeros, but with stimulus onset the first component is set to '1'. With each discrete time step this '1' is shifted to the next component such that component i has the value of '1' if the stimulus onset was exactly $i-1$ time steps ago. The length of the vector determines for how long the stimulus onset can be 'remembered'. The value of a stimulus is estimated with the help of weight vector w that has the same length as the state vector. The weight vector is modified during learning and is used at each discrete time step t to form reward predictions $P(t) = s(t) \cdot w(t)$, where '·' is the dot product.

Changes in the reward prediction in two successive time steps $(P(t-1) - P(t))$ provide an estimate of the reward at time step t, which is $r(t)$. If the estimate is good, the difference between those two should be zero. If the estimate is bad, the difference can be either positive or negative and the estimate should be improved. Thus, this difference $(PE = r(t) - P(t-1) + P(t))$ yields an error in the reward prediction. Commonly, $PE(t)$ is therefore referred to as prediction error and is used to update the weight vector to improve future reward predictions. The weight vector is changed by: $\Delta w = \alpha PE(t)e(t)$, where $0 < \alpha \leq 1$ is a learning-rate and $e(t)$ is a so-called eligibility trace. The eligibility trace contains past stimulus representations that are used for temporal-credit assignment ('what stimulus in the past might have caused the current reward?'). It can be determined recursively, such that $e(t+1) = \lambda e(t) + s(t)$. The parameter $0 \leq \lambda \leq 1$ determines whether only rather recent (low λ) or also more remote (high λ) events are considered responsible for current rewards.

The actor component also uses the prediction error to learn which actions lead to rewards. Each action a is associated with a scalar weight w_a which are updated similar to the stimulus weight vectors: $\Delta w_a(t) = \beta PE(t)e_a(t)$. β is the learning-rate for action learning and $e_a(t)$ is the eligibility trace of each action.

A trial consisted of 15 time steps. The stimulus was presented at time step 5. At the same time step an action was selected on the basis of the action weights. If the correct action was selected a reward was given at time step 10. Big rewards had a value of '2', small ones had a value of '1'. Between trials a random inter-trial interval of 20–60 time steps was inserted. Parameter values in the critic were chosen to match DA cell activity in a reward-learning task [25]. Actor component parameters were chosen to fit the time course of the behavioral data reported here. Parameters values were $\alpha = 0.005$, $\lambda = 0.9$, $\beta = 0.025$, and state and weight vector length was 11.

Action selection was implemented with a Boltzman distribution providing a probability to choose action a: $P_a(t) = \exp(\gamma w_a(t)) / \sum_{a' \in A} \exp(\gamma w_{a'}(t))$ with an inverse temperature $\gamma = 1$ [30,8]. The set of actions A consisted of: peck A, peck B, or do nothing. Initially, the weights for peck A and peck B were set to zero, while 'do nothing' had a small positive weight (0.2). We simulated 50 experiments with small and 50 experiments with big rewards, each consisting of 400 trials. Afterwards, correct responses were assessed as percentages on the basis of 40 consecutive trials. Mean values and standard errors were determined across experiments with the same reward value.

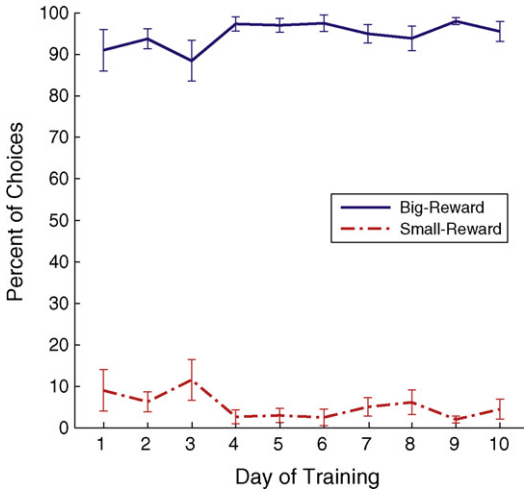

Fig. 2. Preference (mean with standard error) on reward-choice blocks within the color-choice training-sessions. All animals choose the large-reward (solid line) reliably over the small-reward (dashed line). *X*-axis: day of training, *Y*-axis: percent choice of large-reward.

3. Results

3.1. Behavior

Of the 12 animals in training, ten reached criterion on the reward-discrimination (three consecutive days over 80% choice of the big reward) and went on to be tested on the color-discrimination. For these 10 animals, the high level of reward-discrimination was maintained throughout all consecutive sessions (Fig. 2). Training of the remaining two animals was discontinued and they were omitted from analysis.

All animals learned the color-discrimination task within 10 days of training, the criterion (75% correct, big- and small-reward trials combined) was reached after a mean of 4.50 (±1.27 standard deviation) days. The size of the reward had a decisive influence on

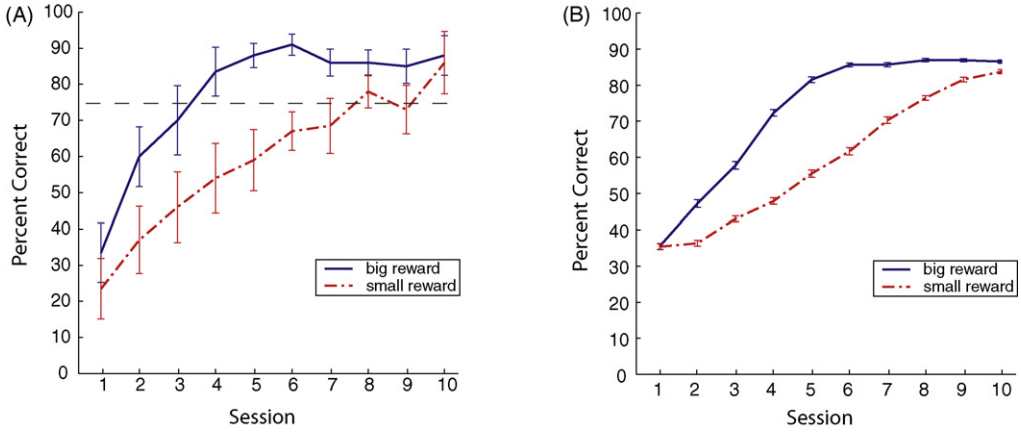

Fig. 3. (A) Acquisition of the color-choice task. Pigeons performance on the first 10 days of training. Percent choice (mean with standard error) of the S++ over the S− is depicted with a solid line, choice of the S+ over the corresponding S− is depicted with a dashed line. The dashed horizontal line represents the criterion of 75% correct. X-axis: training-sessions, corresponding to 40 trials, Y-axis: percent correct trials. (B) Modeling of the color-choice task, depicted are the results of 100 simulations. Percent choice (mean with standard error) of the S++ over the S− is depicted with a solid line, choice of the S+ over the corresponding S− is depicted with a dashed line. *X*-axis: sessions, consisting of 40 learning trials each, *Y*-axis: percent correct trials. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

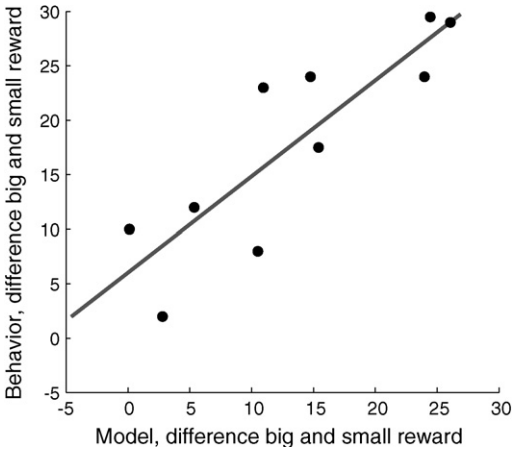

Fig. 4. The model accurately reflects the influence of reward-magnitude on learning. Predictions for the small-reward were subtracted from the predictions for the large-reward. This difference is plotted for the behavioral data (*Y*-axis) and for the model data (*X*-axis).

learning-rates (Fig. 3a). While acquisition of color-pairs reinforced by a large-reward took on average 3.40 (±0.97 standard deviation) days, acquisition of color-pairs reinforced with a small-reward took almost twice as long with 5.9 (±2.33 standard deviation) sessions of training. This difference is significant ($p = 0.0116$). In addition to comparing the number of sessions to criterion, both learning curves were compared directly, this difference was found to be significant ($p = 0.005$). Note that this second test includes the initial and final training-sessions, in other words those sessions in which the animals performed at chance or at maximal performance on both conditions.

3.2. Model

The learning curves of the model resemble the animal behavior (Fig. 3b) with acquisition on the large-reward condition exceeding that on the small-reward condition. To assess if simulated data reflects the difference between the conditions adequately the correlation coefficient between the difference-values for the behavioral- and the modeling data was determined (Fig. 4). The differences show a strong linear correlation ($r = .8614$; $p = 0.0014$) indicating that the model accurately reflects the influence of reward-magnitude on learning-rate.

4. Discussion

The aim of the present study was to test a prediction of reinforcement learning models. These models imply that learning-rates depend on the magnitude of reward delivered after correct responses. To assess this prediction, pigeons were trained on a color-discrimination task with different reward-magnitudes. In line with reinforcement learning models, a large-reward led to fast acquisition of the task, whereas a small-reward led to slow acquisition of the task. As an additional measure, the difference between the acquisition of large- and acquisition of small-rewards was calculated and compared between animal behavior and an implementation of reinforcement learning. Behavior and model were linearly related with respect to this measure. These results imply that TD-models of reinforcement learning accurately predict animal behavior with respect to the influence of reward-magnitude on learning-rates.

We believe that a neural implementation of TD-learning offers a compelling explanation for the observed difference in acquisition. However, motivational influences offer a potential alternative explanation. Various studies have shown that during learning, response-rate or running-speed of animals are modulated by the size of forthcoming rewards [23,19,13,4,7,20]. These results were often interpreted in the light of incentive salience or motivation [2,19]. While we cannot exclude such a motivational interpretation, we believe that TD-learning offers a more parsimonious account for our data. First, the TD-model accurately reflects animal behavior with respect to the difference between large and small-rewards. Importantly, it does so as an intrinsic property of TD-models, without the inclusion of a separate 'motivational module'. Second, if motivation differed greatly between the large-reward and the small-reward color-pairs the animals' performance would reflect this difference also after learning. However, performance reached asymptote on the same level for large- and small-rewards, suggesting that there was no overall effect of motivation on animal performance. Third, we believe that the paradigm used in the present study, forced choice, is far less susceptible to motivational effects than classical paradigms employing response-rate or running-speed in a maze, since these remain sensitive to reward-magnitude after learning. This was already concluded by Crespi who argued that measures of response-strength quantify performance and therefore motivation while error-measures can be used to quantify learning [5,6].

Another line of research on reinforcement magnitude led to the observation that such effects are strongly modulated by subjective experience. Changing the amount of reinforcer received by a single subject for responding, say from a large to a small-reward will result in a large deterioration of performance. If, on the other hand, different subjects are reinforced with different amounts of reinforcer the effect will be a lot less pronounced [3,5,6]. In line with these results, it has been shown that the responses of single neurons involved in reward-processing, are not merely tuned to absolute, physical properties of reinforcement. These neurons rather respond to subjective value of reward, scaled to other available rewards [39]. Consequently, we chose to use a within-subject design to induce a subjective difference in the perception of reinforcement as is implied in the TD-model.

The neural basis of reward-based learning has been an active area of research for several decades. To date there is consensus that the basal ganglia along with midbrain dopaminergic neurons and their thalamo-cortical target areas lie at the heart of reward processing and of reward-based learning [32,34,41,10,11]. Schultz et al. reported in 1997 [31] that single dopaminergic neurons in the midbrain of primates are activated in accordance with a TD prediction error. These results have later been replicated in various studies [35] and it is now widely accepted that dopaminergic neurons carry a prediction error signal [for a different perspective: 2]. This signal finds one of its uses in the striatum to aid learning related processes. Release of dopamine in the striatum can be observed after the presentation of a contingent CS, but not after a non-contingent CS [16]; learning-rate can be increased by microstimulation in the dorsal striatum during the reinforcement-period of a visuo-motor association task [42]; during learning, the activation of striatal neurons precedes that of prefrontal neurons [27]; and dopamine mediates plasticity in cortico-striatal circuits [28,41].

Tobler et al. [39] showed that the responses of dopaminergic neurons are sensitive to the magnitude of forthcoming rewards. Thus, at the neural level of dopamine neurons reward-magnitude is encoded, as required for a neural implementation of reinforcement learning. However, it is unknown how this information about reward-magnitude is read-out at target structures, such as the striatum. To effectively modulate learning-rate, striatal dopamine receptors should show concentration-specific effects which allow the manifestation of different learning-rates in the striatum or downstream targets. Further studies of different dopamine receptor

subtypes might provide interesting insights on their involvement in reward-magnitude modulated task acquisition.

We believe that the paradigm presented here is a useful tool to further investigate this issue. In this paradigm, discrimination of rewards and the influence of reward-magnitude on learning can be assessed by distinct behavioral measures, the choice of large- over small-reward on one hand and the acquisition of the color-discrimination on the other hand. This distinction offers the possibility to pit the discrimination of different reward-magnitudes against the influence of reward-magnitude on learning-rate. Hence, it is a tool to investigate the neural structures and pharmacological substrates of a reward modulation of learning.

In the future we hope to elucidate how, in this learning regime, the contrast between different reward-magnitudes is generated; is learning to predict large-rewards fostered, learning to predict small-rewards hindered or do both mechanisms interact; what is the role of different dopamine-receptors and of striatal regions in discriminating reward-magnitudes and learning from different rewards.

Acknowledgement

This work was supported by the BMBF grant 'reward-based learning'.

References

[1] Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 2005;47(1):129–41.
[2] Berridge KC. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 2007;191:391–431.
[3] Black RW. Shifts in magnitude of reward and contrast effects in instrumental and selective learning: a reinterpretation. Psychol Rev 1968;75(2):114–26.
[4] Collier G, Marx MH. Changes in performance as a function of shifts in the magnitude of reinforcement. J Exp Psychol 1959;57(5):305–9.
[5] Crespi LP. Quantitative variation in incentive and performance in the white rat. Am J Psychol 1942;55:467–517.
[6] Crespi LP. Amount of reinforcement and level of performance. Psychol Rev 1944;51:341–57.
[7] Denny MR, King GF. Differential response learning on the basis of differential size of reward. J Genet Psychol 1955;87:317–20.
[8] Doya K. Modulators of decision making. Nat Neurosci 2008;11(4):410–6.
[9] Dufort RH, Kimble GA. Changes in response strength with changes in the amount of reinforcement. J Exp Psychol 1956;51(3):185–91.
[10] Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci 2005;8(11):1481–9.
[11] Graybiel AM. The basal ganglia: learning new tricks and loving it. Curr Opin Neurobiol 2005;15:638–44.
[12] Güntürkün O. Avian and mammalian "prefrontal cortices": limited degrees of freedom in the evolution of the neural mechanisms of goal-state maintenance. Brain Res Bull 2005;66:311–6.
[13] Guttman N. Operant conditioning, extinction, and periodic reinforcement in relation to concentration of sucrose used as reinforcing agent. J Exp Psychol 1953;46(4):213–24.
[14] Guttman N. Equal-reinforcement values for sucrose and glucose solutions compared with equal-sweetness values. J Comp Physiol Psychol 1954;47(5):358–61.
[15] Hutt PJ. Rate of bar pressing as a function of quality and quantity of food reward. J Comp Physiol Psychol 1954;47(3):235–9.
[16] Ito R, Dalley JW, Robbins TW, Everitt BJ. Dopamine release in the dorsal striatum during cocaine-seeking behavior under the control of a drug-associated cue. J Neurosci 2002;22(14):6247–53.
[17] Jenkins WO, Clayton FC. Rate of responding and amount of reinforcement. J Comp Physiol Psychol 1949;42(3):174–81.
[18] Kalenscher T, Windmann S, Diekamp B, Rose J, Güntürkün O, Colombo M. Single units in the pigeon brain integrate reward amount and time-to-reward in an impulsive choice task. Curr Biol 2005;15(7):594–602.
[19] Kendler HH, Kimm J. Reversal learning as a function of the size of the reward during acquisition and reversal. J Exp Psychol 1967;73(1):66–71.
[20] Maher WB, Wickens DD. Effect of differential quantity of reward on acquisition and performance of a maze habit. J Comp Physiol Psychol 1954;47(1):44–6.
[21] Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive hebbian learning. J Neurosci 1996;16(5):1936–47.
[22] Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 2004;43(1):133–43.
[23] Neuringer AJ. Effects of reinforcement magnitude on choice an rate of responding. J Exp Anal Behav 1967;10:417–24.
[24] Nissen HW, Elder JH. The influence of amount of incentive on delayed response performance of chimpanzees. J Genet Psychol 1935;47(1):49–72.
[25] Pan WX, Schmidt R, Wickens JR, Hyland BI. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci 2005;25(26):6235–42.
[26] Pan W-X, Schmidt R, Wickens JR, Hyland BI. Tripartite mechanism of extinction suggested by dopamine neuron activity and temporal difference model. J Neurosci 2008;28(39):9619–31.
[27] Pasupathy A, Miller EK. Different time courses of learning-related activity in the prefrontal cortex and striatum. Nature 2005;433:873–6.
[28] Reynolds JN, Hyland BI, Wickens JR. A cellular mechanism of reward-related learning. Nature 2001;413(6851):67–70.
[29] Rose J, Otto T, Dittrich L. The biopsychology-toolbox: a free, open source Matlab-toolbox for the control of behavioral experiments. J Neurosci Methods 2008;175:104–7.
[30] Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science 2005;310(5752):1337–40.
[31] Schultz W, Dayan P, Montague R. A neural substrate of prediction and reward. Science 1997;275:1593–9.
[32] Schultz W. Multiple reward signals in the brain. Nat Rev Neurosci 2000;1:199–207.
[33] Schultz W. Getting formal with dopamine and reward. Neuron 2002;36:241–63.
[34] Schultz W. Behavioral dopamine signals. Trends Neurosci 2007;30(5):203–10.
[35] Schultz W. Multiple dopamine functions at different time courses. Annu Rev Neurosci 2007;30:259–88.
[36] Sutton RS, Barto AG. Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 1981;88(2):135–70.
[37] Sutton RS. Learning to predict by the methods of temporal differences. Machine Learn 1988;3:9–44.
[38] Thompson RH, Ménard A, Pombal M, Grillner S. Forebrain dopamine depletion impairs motor behavior in lamprey. Eur J Neurosci 2008;27(6):1452–60.
[39] Tobler PN, Fiorillo CD, Schultz W. Adaptive coding of reward value by dopamine neurons. Science 2005;307:1642–5.
[40] Waelti P, Dickinson A, Schultz W. Dopamine responses comply with basic assumptions of formal learning theory. Nature 2001;412(6842):43–8.
[41] Wickens JR, Horvitz JC, Costa RM, Killcross S. Dopaminergic mechanisms in actions and habits. J Neurosci 2007;27(31):8181–3.
[42] Williams ZM, Eskandar EN. Selective enhancement of associative learning by microstimulation of the anterior caudate. Nat Neurosci 2006;9(4):562–8.
[43] Zeaman D. Response latency as a function of the amount of reinforcement. J Exp Psychol 1949;39(4):466–83.