EJN fens EUROPEAN JOURNAL OF NEUROSCIENCE

European Journal of Neuroscience, Vol. 28, pp. 1849–1866, 2008 doi:10.1111/j.1460-9568.2008.06480.x

BEHAVIORAL NEUROSCIENCE

Anticipatory reward signals in ventral striatal neurons of behaving rats

Mehdi Khamassi,[1,2,*] Antonius B. Mulder,[1,*,†] Eiichi Tabuchi,[1,‡] Vincent Douchamps[1] and Sidney I. Wiener[1]

[1]Laboratoire de Physiologie de la Perception et de l'Action, Collège de France, CNRS, 11 pl. Marcelin Berthelot, 75231 Paris Cedex 05, France

[2]ISIR, Université Pierre et Marie Curie – Paris 6, 75016 Paris, France

Keywords: accumbens, Actor–Critic, reinforcement learning, striatum, TD learning

Abstract

It has been proposed that the striatum plays a crucial role in learning to select appropriate actions, optimizing rewards according to the principles of 'Actor–Critic' models of trial-and-error learning. The ventral striatum (VS), as Critic, would employ a temporal difference (TD) learning algorithm to predict rewards and drive dopaminergic neurons. This study examined this model's adequacy for VS responses to multiple rewards in rats. The respective arms of a plus-maze provided rewards of varying magnitudes; multiple rewards were provided at 1-s intervals while the rat stood still. Neurons discharged phasically prior to each reward, during both initial approach and immobile waiting, demonstrating that this signal is predictive and not simply motor-related. In different neurons, responses could be greater for early, middle or late droplets in the sequence. Strikingly, this activity often reappeared after the final reward, as if in anticipation of yet another. In contrast, previous TD learning models show decremental reward-prediction profiles during reward consumption due to a temporal-order signal introduced to reproduce accurate timing in dopaminergic reward-prediction error signals. To resolve this inconsistency in a biologically plausible manner, we adapted the TD learning model such that input information is nonhomogeneously distributed among different neurons. By suppressing reward temporal-order signals and varying richness of spatial and visual input information, the model reproduced the experimental data. This validates the feasibility of a TD-learning architecture where different groups of neurons participate in solving the task based on varied input information.

Introduction

Prefrontal cortex–basal ganglia loops have been identified as instrumental in orchestrating behavior by linking past events and anticipating future events (Alexander *et al.*, 1990; Fuster, 1997; Samejima & Doya, 2007). Within these loops, it is proposed that the striatum enables learning mechanisms for organizing action sequences, particularly those with time scales orders of magnitude greater than those of postsynaptic events (Schultz *et al.*, 1997; Graybiel, 1998; Hikosaka *et al.*, 1999). Indeed, some striatal neurons are selectively active in the successive actions comprising goal-directed behaviors (Itoh *et al.*, 2003; Mulder *et al.*, 2004; Schmitzer-Torbert & Redish, 2004), and yet others fire in relation to (sometimes anticipating) rewards including food, drink, habit-forming drugs and intracranial electrical stimulations (Hikosaka *et al.*, 1989; Schultz *et al.*, 1992; Wiener, 1993; Lavoie & Mizumori, 1994; Miyazaki *et al.*, 1998; Shibata *et al.*, 2001; Daw *et al.*, 2002; Setlow *et al.*, 2003; Nicola *et al.*, 2004; Wilson & Bowman, 2005).

The dichotomy between striatal neurons representing actions and striatal neurons representing rewards led to the design of Actor–Critic models of this brain area (Houk *et al.*, 1995; see Joel *et al.*, 2002; Khamassi *et al.*, 2005 for reviews). In such models, the Actor is a 'memory structure' which is responsible for selecting actions, whereas the Critic evaluates the actions made by the Actor (Sutton & Barto, 1998). More precisely, the Critic learns to predict reward, and the mismatch between consecutive reward-predictions and actual reward occurrences is used as a temporal difference (TD) reinforcement signal to update the Actor.

The existence of strong projections from the ventral striatum to the dopaminergic ventral tegmental area (VTA) and substantia nigra pars compacta (SNc; Haber *et al.*, 2000; Joel & Weiner, 2000; Thierry *et al.*, 2000), and the discovery of the TD-like reward-prediction error responses of these areas (Schultz *et al.*, 1997), support the hypothesis that the ventral striatum could play the role of a Critic (O'Doherty *et al.*, 2004). However, the suitability of ventral striatal signals for elaborating such TD learning processes remains to be established.

In order to explain dopaminergic neurons' depression in activity when an expected reward is omitted (Schultz *et al.*, 1997), the Montague *et al.*

Correspondence: Dr S. I. Wiener, as above.
E-mail: sidney.wiener@college-de-france.fr

*M.K. and A.B.M. contributed equally to this work.

[†] *Present address*: Cognitive Neurophysiology-CNCR, Department of Anatomy and Neurosciences, VU University Medical Center, van de Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands.

[‡] *Present address*: Department of Analysis of Brain Function, Faculty of Food Nutrition, Toyama College, 444 Gankaiji, Toyama 930-0193, Japan.

Received 12 June 2008, revised 12 August 2008, accepted 2 September 2008

(1996) and Suri & Schultz (2001) models propose that the striatum processes a time-based representation of stimuli. This 'complete serial compound stimulus' component provides the Critic with temporal-order information and enables it to 'count' the time bins between a conditioned stimulus and a reward. One of the consequences of this component is a linear decrease in amplitude of reward-prediction activity during the reward delivery period, so that it becomes null at the end of reward delivery (see Suri & Schultz, 2001; Figs 3–5).

To test this prediction, we recorded ventral striatal neurons in rats as they approached goals in a plus maze and then stood still awaiting successive rewards presented at 1-s intervals. The TD-learning model was then adapted to reproduce the neural activity recorded.

Materials and methods

Animals and apparatus

Seven Long–Evans male adult rats (220–240 g) were obtained (from the Centre d'Elevage René Janvier, Le Genest-St-Isle, France) and kept in clear plastic cages bedded with wood shavings. The rats were housed in pairs while habituating to the animal facility environment. They were weighed and handled each work day. Prior to training they were placed in separate cages and access to water was restricted to maintain body weight at not less than 85% of normal values (as calculated for animals of the same age provided *ad libitum* food and water). The rats were examined daily for their state of health and were rehydrated at the end of each work week. This level of dehydration was necessary to motivate performance in the behavioral tasks, and the rats showed neither obvious signs of distress (excessive or insufficient grooming, hyper- or hypoactivity, or aggressiveness) nor health problems. The rats were kept in an approved (City of Paris Veterinary Services) animal care facility in accordance with institutional (CNRS Comité Opérationnel pour l'Ethique dans les Sciences de la Vie), national (French Ministère de l'Agriculture, de la Pêche et de l'Alimentation No. 7186) and international (US National Research Council Guide for the Care and Use of Laboratory Animals, 1996) guidelines. A 12–12 h light–dark cycle was applied.

Training and experiments took place in a four-arm plus maze. The arms were 70 cm long and 30 cm wide with 40 cm high sloped black walls while the center was a 30 × 30 cm square. This was placed in a darkened square room (3 × 3 m) bordered by opaque black curtains (Fig. 1). At the end of each of the four arms was an alcove (30 × 30 × 30 cm) containing a water reservoir and a large, highly contrasted, three-dimensional visual cue. The cues were identical in each of the boxes but could be illuminated independently. Room cues included a wide inverted-T shaped white poster board (185 × 60 cm) as well as a white rectangular box (56 × 25 cm), each mounted 70 cm from the platform on walls respectively opposite or adjacent to the

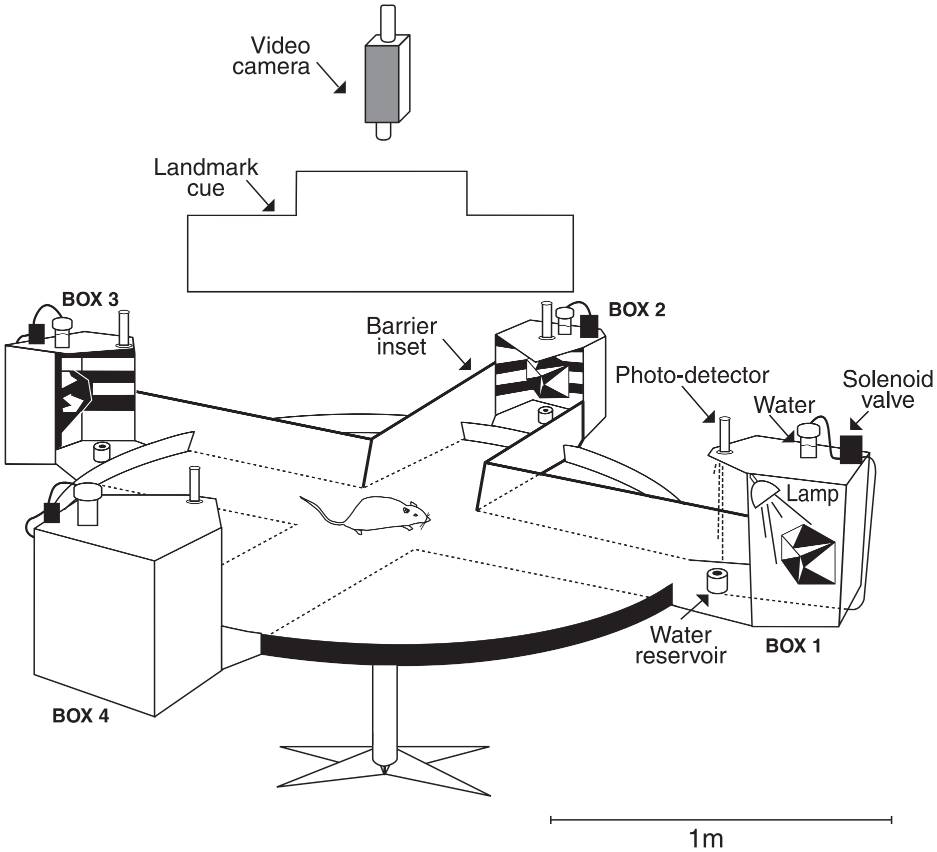


FIG. 1. The experimental apparatus. The rat performed the behavioral task on a 180-cm-diameter platform with a low border. Barriers placed on the platform (broken lines) restricted the movements of the rats to four alleys. Four reward boxes (30 × 30 × 30 cm) were attached to the edge of the platform and were equally spaced and oriented toward the corners of the experimental room. Each box contained identical, highly contrasted polyhedrons suspended in front of a striped background. Each reward box could be illuminated independently under computer control. The main sources of illumination in the experimental room were the lamps directed towards the salient cues in the reward boxes, the overhead lamp and two miniature lamps on the headstage of the rat (Adapted from Tabuchi *et al.*, 2000).

entrance of the curtained area. The poster board was spotlit by a ceiling-mounted incandescent lamp (60 W) during both training and recording sessions.

Each reward box was equipped with automated water delivery and infrared photo-emitter and detector systems. At the entry of each reward alcove stood a short (3 cm high) cylindrical block (the 'water reservoir'). Tubing transported water from elevated bottles to computer-controlled solenoid valves that in turn led to each water reservoir. When the rat arrived at the water reservoir and blocked the photobeam, the computer triggered release of the water reward(s) there. The volume of the water droplets was calibrated to 30 μL by regulating the time that the solenoid valves remained open. Multiple droplets of water were provided at 1-s intervals. The solenoid valves made an audible click when opening and closing. The times of the photobeam occlusions as well as solenoid valve openings were recorded as event flags in the data file. Photodetectors also registered when the rat arrived at the center of the maze.

The differentially rewarded plus-maze task

Details of the task and training protocols may be found in Tabuchi *et al.* (2000, 2003) and in Fig. 2. In each session the rats were trained with a novel distribution of different reward volumes at the four respective arms of the maze and then were required to recall the sequence in order of decreasing volume. After this, the reward distribution was changed and a second series of training and recall trials were run while the same cells were recorded.

In the training phase, reward availability was successively signaled by cue lamps in each of the reward boxes. The rat thus went to the respective boxes that provided 7, 5, 3 and then 1 droplets of water. Rats performed this sequence of visits 6–8 times to stimulate learning the amount of water associated with each maze arm. For the multiple rewards at each box, the successive droplets of water were delivered at 1-s intervals while the cue lamp remained lit. After the rat consumed the water it returned to the center of the maze and the lamp on the next arm was then lit automatically.

In the recall phase, all reward alcoves were illuminated at the beginning of each trial and the lights were turned off successively as the rats visited them in order of descending reward value. The task design exploited the tendency for rats to prefer locations with greater rewards (e.g., Brown & Bowman, 1995; Albertin *et al.*, 2000). If the rat entered an arm out of sequence, all cue lamps were turned off and the same lamps were lit again when the rat returned to the maze center. The rats only very rarely continued to the end of the arm in these cases, and thus there was insufficient data to analyze error trials (and the goal here was only to analyze the dynamics of reward-anticipatory responses).

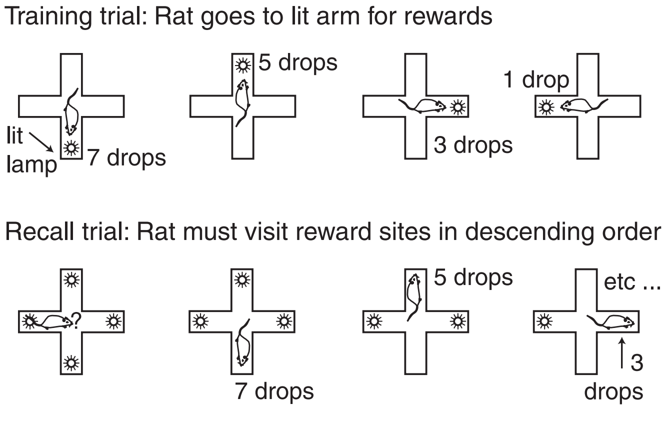

Training trial: Rat goes to lit arm for rewards
lit
lamp
7 drops
5 drops
3 drops
1 drop
Recall trial: Rat must visit reward sites in descending order
?
7 drops
5 drops
etc ...
3
drops

FIG. 2. The experimental task. First the rats performed a series of training trials where the correct choice was guided by the lit cue lamp in the appropriate reward box. Each trial comprised a sequence of visits to the four reward boxes providing 7, 5, 3 and 1 droplets of water. During recall trials all cue lamps were lit, then were turned off one by one as the rat visited the reward boxes in the same order of descending reward value. Reward values were then reassigned for the second half of the session, and were also changed daily (Adapted from Tabuchi et al., 2000).

Electrode implantation and recordings

Electrodes were surgically implanted after the performance level exceeded 70% correct (rewarded) visits (usually after 4–6 weeks of training). The rat was returned to *ad libitum* water, tranquillized with 0.1 mL of 2% xylazine (i.m.) and anesthetized with 40 mg/kg pentobarbital intraperitoneally. Two bundles of eight 25-μm formvar-insulated nichrome wires with gold-plated tips (impedance 200–500 kΩ) were stereotaxically implanted. Each bundle was installed in a guide tube (a 30-gauge stainless steel cannula) and mounted on one of two independently advanceable assemblies on a single headstage (Wiener, 1993). A ground screw was installed in the cranial bone. One group of electrodes was placed above either the ventrolateral shell region of the nucleus accumbens (Acb) (AP 10.7–11.2, ML 1.7–2.2) or the medial shell of Acb (AP 11.2–11.6, ML 0.7–0.9). While the ventral striatum was the target here, occasional neurons with reward-related responses observed in adjacent structures such as ventral caudate are also reported. The second bundle was placed above the hippocampus (data reported in Tabuchi *et al.*, 2000, 2003). About 1 week later, after complete recovery from the surgery, water restriction and training were resumed. The screws of the advanceable electrode drivers were gradually rotated daily until neurons were isolated (the drivers advanced 400 μm for each full rotation); then multiple single units were recorded as the rat performed the tasks. The electrodes were advanced at least 3 h prior to recording sessions to promote stability.

Electrode signals passed through FETs (field-effect transistors), then were differentially amplified (10 000×) and filtered (300 Hz to 5 kHz, notch at 50 Hz). Single-unit activity was discriminated *post hoc* with Datawave® software, where single-unit isolation was performed using eight waveform parameters (positive, negative and entire spike amplitude, spike duration, amplitude windows immediately prior to and after the initial negative-going peak, and time until maxima of positive and negative peaks) on the filtered waveform signals. Isolation was confirmed in interspike interval histograms which had, on average, only 0.3% occupancy of the first three 1-ms bins corresponding to the refractory period. Waveforms are presented in Supporting information, Fig. S1 and as insets to raster and histogram figures.

Two small lamps (10 cm separation) were mounted 10 cm above the headstage. Reflectors were attached to the rostral lamp to aid the tracking system in distinguishing it from the caudal lamp. The two lamps were detected with a video camera mounted above the platform and transmitted to a video tracking system (DataWave, Longmont, CO, USA) and a video monitor. All of the action potential (digitized waveforms and timing) and behavioral (position of the animal, photobeam crossings and water delivery) data were simultaneously acquired on a personal computer with software operating under DOS (DataWave, Longmont, CO, USA).

In preparation for recording sessions, the rat was placed in a cage with transparent plastic walls (and no wood shavings) then brought into the experimental room. The recording cable was attached to the headstage and the rat was placed in a cubic cardboard box (with sides ~40 cm). Then the electrode recording channels were examined for signs of discriminable neuronal activity. If this was successful, the data acquisition system was initialized and the lamp assembly was attached. The rat was then placed in the experimental apparatus where the lamp at the first reward box was already lit. No attempts were made to disorient the rat, and the lengthy training period assured that the environment was familiar. The rats always immediately started performing the task. Sessions usually lasted ~20 min.

Data analysis

Data from all recorded neurons with average firing rates > 0.1 impulses/s during the experiment were submitted to statistical analyses (11 neurons were excluded by this criterion). The synchronization point for analyses of cell activity was selected as the instant that the computer triggered the first droplet of water after the tip of the rat's muzzle blocked the photobeam at the reward boxes. In this experimental design, ANOVA was selected for determining the correlations of firing rate of the neurons with spatial position, behavior and task phase. In order to better approximate a gaussian distribution, spike count data were first transformed (the sum of the square root of the spike count was summed with the square root of the spike count incremented by one; Winer, 1971). ANOVA has been shown to be robust even in cases where the underlying distribution is not perfectly gaussian (Lindman, 1974).

Two different analyses of ANOVAs tested for the first two or all of the following three factors: (i) behavioral correlates: comparisons of firing rates during reward site approach, arrival and water consumption (two 0.5-s periods prior to and after delivery of the first droplet of water); (ii) position correlates: differences in firing rate when the rat occupied the different maze arms; and (iii) comparisons between phases of the experiment (training vs. recall phases and after changes in the reward distribution). Data were also recombined from recordings on different arms that provided the same reward volume during the course of a session. Statistical results were considered significant at $P < 0.05$. The Student–Newman–Keuls test was employed for *post hoc* analyses. The above ANOVAs and *post hoc* tests were performed with STATISTICA® (Statsoft, Tulsa, OK, USA). A one-way ANOVA compared response amplitudes among successive drops of water (measured in 1-s intervals between mean activity minima) with the Tukey–Kramer *post hoc* test (Matlab®). The Pearson correlation test (Matlab®) compared cells' activity profiles and the computational model's profiles, where neural activity of individual cells was averaged over trials in 250-ms bins during the reward consumption period at the four maze boxes [windows of $(10s + 8s + 6s + 4s) \times 4 = 112$ values]. The chi-squared test (Matlab®) was used for testing for difference in distributions of respective cell response types among different striatal regions. Other tests were performed with Microsoft Excel®.

Histology

After experiments were completed the rat was rehydrated for at least a day, then deeply anesthetized with pentobarbital. A small electrolytic lesion was made by passing DC current (20 μA, 10 s) through one of the recording electrodes to mark the location of the electrode tip. Intracardial perfusion with saline was followed by 10% formalin in 0.1 M phosphate buffer (pH 7.4). Serial frozen sections (50 μm thickness) were stained with Cresyl violet. Recording sites were reconstructed by detecting the small electrolytic lesion and the track left by the guide tube, then taking into account the distance that the microelectrode driver had been advanced from the point of stereotaxic placement of the electrodes. The recording sites were calculated by interpolating along the electrode track between the lesion site and the implantation site.

Results

Task performance levels

In eight rats recordings were made in 35 experimental sessions. In all cases performance was nearly perfect on light-cued training trials. As withholding rewards can provoke rats toward disruptive behavior at reward sites, erroneous arm entries were signalled by turning off lights and rats often did not continue on to the end of the arm. In the sessions described here, the mean percentage of correct visits was 80 ± 11%, ranging from 61 to 100%. The number of completely correct trials, that is, four visits in sequence of descending reward volume, was 36 ± 32% (mean ± SEM) and ranged from 0 to 100% in individual sessions. (Note that the probability of correctly performing a complete trial by chance is ~4%, that is $0.25 \times 0.33 \times 0.50$).

Cell localization

Electrode placements were intentionally made in different parts of the ventral striatum in order to explore diverse subregions for possible reward-associated responses. Figure 3 shows that recording sites were distributed in the core of the nucleus accumbens, the medial shell of the nucleus accumbens and the ventromedial part of the caudate nucleus. There was no anatomical segregation of different response types (χ^2, $P > 0.05$).

Cell activity profiles

The ANOVAs revealed significant behavioral correlates in ~75% of the neurons recorded in the nucleus accumbens core (33 of 43), the accumbens shell (60 of 81) and the ventromedial part of the caudate nucleus (53 of 68). The present study focuses only on those cells that showed significant changes in firing rate when rewards were delivered ($n = 46$; other neurons reported in Mulder *et al.*, 2004 are discussed below).

Among these 46 cells showing reward-related activity, we distinguishd phasically-firing neurons (PFNs) and tonically-firing neurons (TFNs), following the terminology adopted by Schmitzer-Torbert & Redish (2004) for rats. This is because it is not clear whether the distinction between phasic vs. tonic neurons (TANs, tonically active neurons) observed in primates (Apicella *et al.*, 1996) can be applied in rats. The TFN group can include putative TANs (as will be seen in Fig. 9), but other neurons in this group do not have qualifying properties such as a very low firing rate, and thus may not be cholinergic neurons (Graybiel & Kimura, 1995). As in previous work (Mulder *et al.*, 2005) TFNs were identified principally by (i) the absence of 'silent' periods (when the firing rate was <1 impulse/s) of 2 s or longer along the course of a trial, and (ii) a significant decrease or increase firing (relative to baseline) during a task event. In contrast, PFNs had silent periods interspersed with brief bouts of behaviorally correlated activity. This pattern of phasic activity superimposed upon negligible background activity is consistent with identification as a medium spiny principal neuron (see Mulder *et al.*, 2005). While only 14 of the total 66 (21%) PFNs with significant behavioral correlates fired during reward delivery, 32 of the 80 (41%) TFNs with behavioral

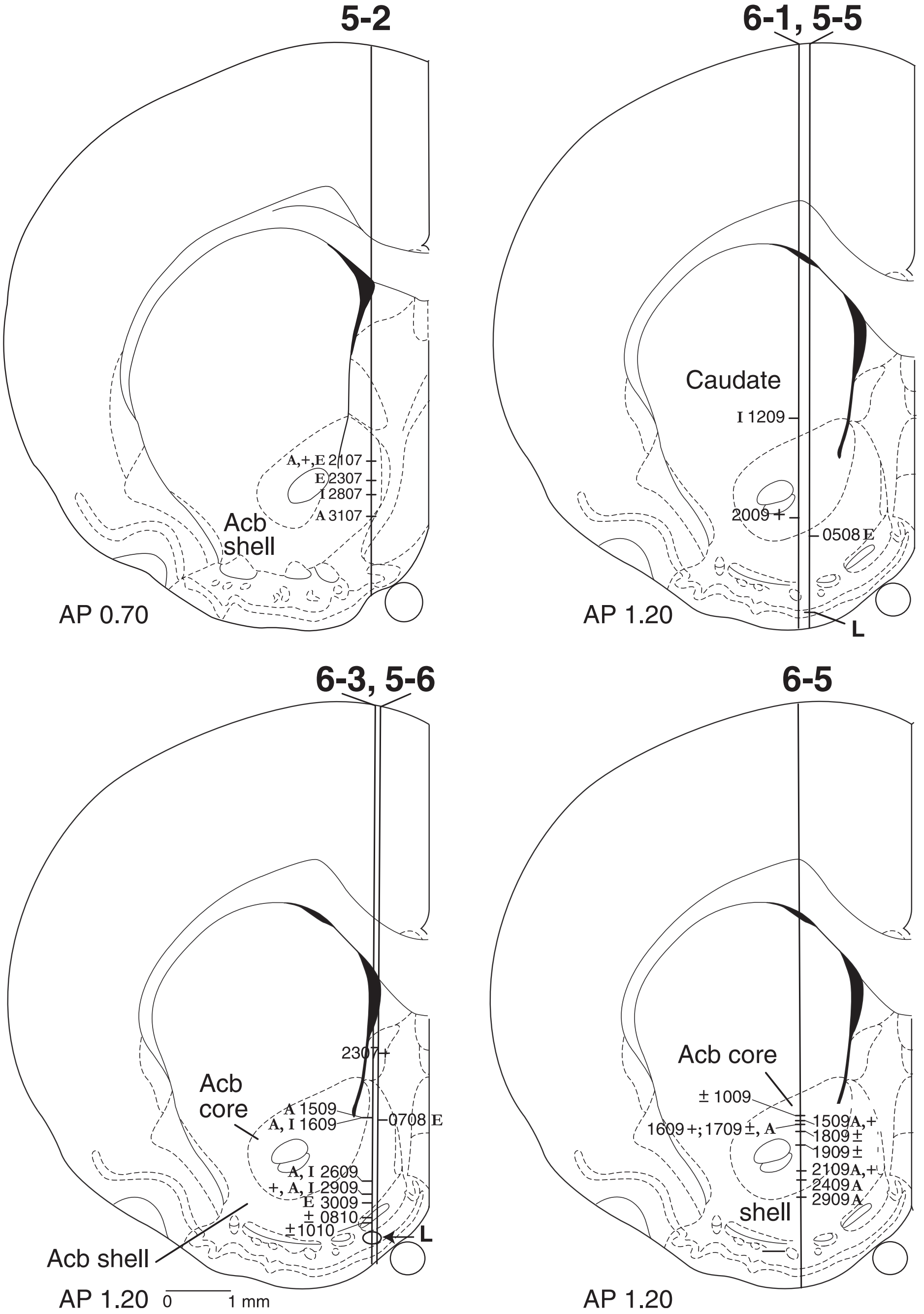


FIG. 3. Reconstruction of recording sites on the basis of histological preparations. Animal identification numbers appear above respective electrode tracks. Recording sites are marked by cross bars and numbers. Neurons are identified according to the following code: A, anticipatory responses for individual droplets of water; E, uniform increase in firing rate during drinking; I, inhibition during drinking; +, excitatory response for first droplet only; ±, Excitation and inhibition during first droplet; L, lesion site. Multiple single neurons recorded at the same site are separated by commas. Histological analyses showed tracks in animal 6–2 were indeed in ventral striatum but sites could not be reconstructed with precision (data not shown). (Figure templates adapted from Paxinos and Watson, 1998, with permission).

correlates had these properties. No other behavioral correlates were observed in these neurons.

Overview of cell response types

Three principal categories of reward-related responses were distinguished to classify the cell activity profiles (Table 1). In the first group, there was a significant phasic firing rate increase prior to and during delivery of the successive droplets of water (n = 14). The second group showed a firing rate increase (n = 14) or mixed excitation and inhibition (n = 7) during delivery of only the first droplet of water. The latter responses do not anticipate later rewards at the same site and thus may be more closely correlated with reward-approach behaviors. Finally, there was a group of neurons with tonic firing rate increases (n = 5) or decreases (n = 6) throughout the period when multiple droplets of water were delivered. Note that some of these behavioral correlates could easily be confounded with one another if recorded in experimental protocols providing only single rewards. Examples and

TABLE 1. List of recorded cells with recording site and cell type categorized by type of correlated activity

Session	Cell no.	P or T	Anatomy
Peridrop excitation, all drops			
522107	02	T	MSh
523107	01	P	MSh
542307	01	T	(unclear)
540108	01	P	MSh
621609	22	T	Lateral core
631509	11	T	MSh
631609	11	T	MSh
632609	41	T	MSh
632909	32	P	VMSh
651509	21	P	Core
651709	01	P	Core
652109	02	T	Core
652409	01	P	Core
652909	01	P	Vsh
First drop, excitatory response			
522107	21	T	MSh
562307	21	T	Septum
612009	11	T	Core
620809	11	T	VMC
620909	12	P	VMC
620909	21	P	VMC
620110a	02	P	Lateral core
621109	11	P	VMC
622609	01	P	Lateral core
632909	42	T	VMSh
651509	01	P	Core
651609	11	T	Core
652109	01	P	Core
652109	03	P	Core
First drop excitatory then inhibitory responses			
622209	01	T	Lateral core
631010	01	T	VP
632909	01	T	VMSh
651009	01	T	Core
651709	21	T	Core
651809	31	T	Core
651909	11	T	Core
General increased activity during drinking			
522307	02	T	MSh
522807	21	T	MSh
550508	03	T	Core/MSh
560708	11	T	MSh
633009	23	T	VMSh
General inhibition during drinking period			
522807	01	T	MSh
611209	01	T	VMC
621909	02	T	Lateral core
630810	21	T	VP/ICj/MSh
631609	01	T	MSh
632609	51	T	MSh

Icj, interstitial n. Cajal; P, phasic; T, tonic; MSh, medial shell; VMC, ventromedial caudate; VMSh, ventromedial shell; VSh, ventral shell; VP, ventral pallidum. A natomical location had unclear histological results.

analyses of these response types will be presented first. Then their coherence with predictions of previous models will be evaluated and an adaptation of the Actor–Critic model will be presented to resolve observed inconsistencies.

Reward-anticipatory responses

Fourteen neurons showed a firing rate increase prior to each successive droplet of water. This anticipatory activity occurred whether the animal was running (prior to the first droplet) or was immobile and waiting for subsequent droplets. Although the activity preceding the first drop of water could be associated with sensory or motor events (the looming image of the lit cue in the reward box, deceleration, assuming an immobile stance), this explanation is not plausible for the responses for the subsequent droplets as the rats invariably remained stably positioned at the water trough. Thus this activity is independent of locomotor behavior. (Simple motor correlates have not been reported this ventral in the striatum; *e.g.*, Shibata *et al.*, 2001; Mulder *et al.*, 2004). The activity was not associated with licking as this started after the solenoid valve clicked and water was released (not shown). Hence, the time course of licking was distinct from the neuron activity profile. Figure 4 shows an example of such activity in a PFN. This nucleus accumbens core neuron started to discharge above baseline 600–800 ms prior to each reward release, with peak activity on average 100 ms before each droplet. Each peak was significantly higher than the baseline activity computed between 2 s and 1 s before the first droplet reward (one-way ANOVA, $F_{8,288} = 17.94$, $P < 0.00001$; Student–Newman–Keuls *post hoc* test). The greatest responses occurred for the first and last drops of water (*post hoc* analyses for paired comparisons, $P < 0.05$). Interestingly, this neuron fired again in the same time window 1 s after the final droplet was delivered. This is consistent with a prediction of yet another reward that was not provided. This anticipatory activity occurred on both visually-guided training trials and memory-guided recall trials (data from both are shown in the figures). This activity is surprising as it occurred after the lamp signaling cue availability had been turned off. Recall that in daily training trials the rats reliably used these same lights to locate the current reward site. This indicates that the neuron did not have access to full information concerning the environment, a point that an accurate model must take into account. Note that in the present case this ‘erroneously predictive’ activity occurred on fewer than half of the trials, yielding smaller histogram peaks than observed for the preceding reward (*post hoc* analyses for paired comparisons, $P < 0.05$). The rat did not consistently depart later from the reward site on trials with the anticipatory responses. Thus this erroneously predictive activity at the level of the single neuron is not necessarily indicative of the expectations of the animal. There was no clear correlation between the appearance of this activity on a given trial and whether there were erroneous visits to other arms immediately prior. There was also no relation between the daily performance level of the rat and the incidence of erroneously predictive activity; the latter appeared even in sessions in which the rat made 90% correct visits. Furthermore, this activity always occurred while the animal still blocked the photobeam at the reward trough. Thus it is parsimonious to consider this activity to be associated with signalling the episodic anticipation of another droplet of water rather than motor preparation of the subsequent departure (as movement timing was the same on trials with and without the predictive activity). Activity in these neurons was not correlated with departures (not shown).

Figure 5 demonstrates this type of response in a ventromedial caudate TFN with a higher firing rate. This neuron started to fire above the background rate at 200 ms prior to the first reward trigger and continued until 300–500 ms afterwards. Similar to Fig. 4, maximal responses occurred at the first reward (one-way ANOVA, $F_{8,216} = 15.96$, $P < 0.01$; *post hoc* test for paired comparisons, $P < 0.05$). However, in contrast to Fig. 4, the peak for the erroneous reward prediction at the end was not significantly smaller than previous peaks (*post hoc* test for paired comparisons, $P > 0.05$). The timing of this final response resembles the preceding ones. The persistence of this activity in the 300–500 ms following the reward is independent of the presence or absence of reward. While not

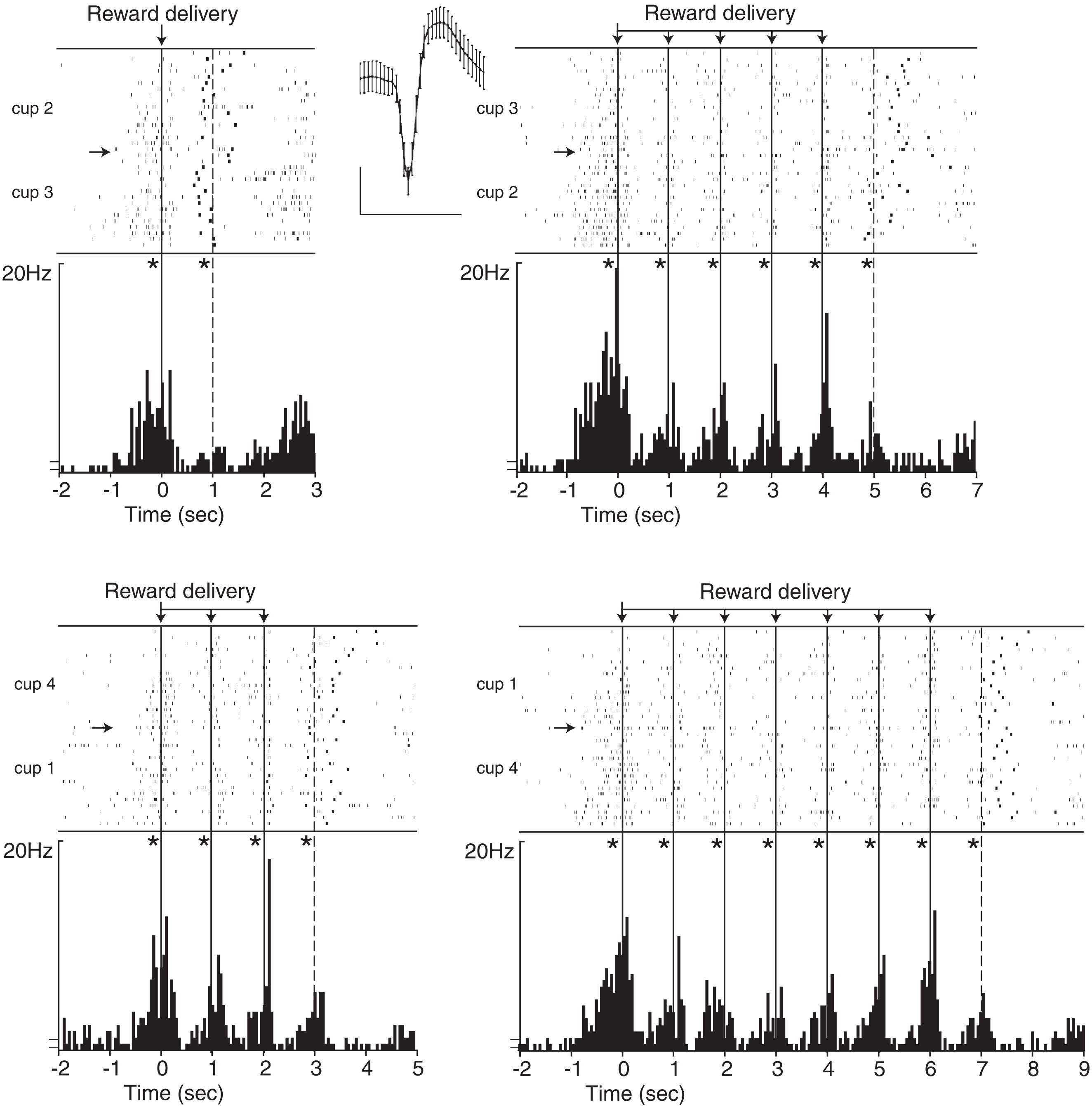
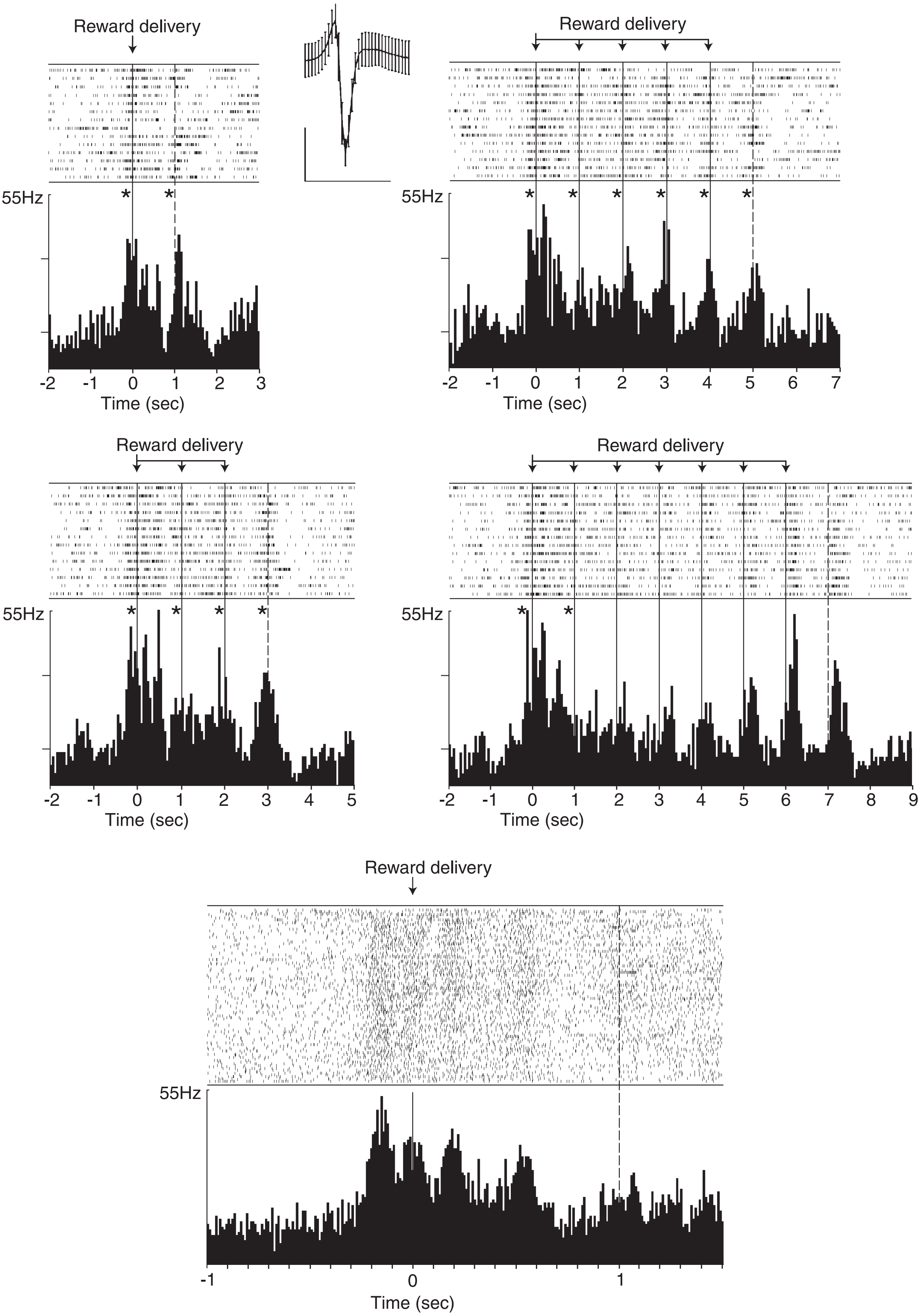


FIG. 4. Phasic excitatory activity predicting reward delivery in a nucleus accumbens core neuron. Raster displays and corresponding histograms (50 ms binwidth) are synchronized with the onset of reward delivery (arrows above). As the reward value distribution was changed in the middle of the session, data have been regrouped from the cups to combine data corresponding to 1, 3, 5 and 7 droplets of water respectively. Arrows at the left in the raster displays separate data acquired at the respective cups. The discharge activity began as early as 800 ms prior to the reward delivery. Stars above histograms indicate peaks significantly higher than the baseline activity. Note that in the lower left panel there is a fourth peak in the histogram at time 3 s, even though no fourth reward was delivered then. The same inaccurate predictive activity also appears in the right panels corresponding to 5 and 7 droplets. Filled squares at the right of the raster displays indicate the animal's departure from the water reservoir. Note that departures are not consistently earlier for trials with no erroneously predictive activity (best visible in lower right panel). Activity at the right border of the panels corresponds to arrivals at the next reward site. At lower left, lower horizontal line indicates mean firing rate from 2 to 1 s prior to reward trigger. Upper horizontal line indicates three times this value. Waveform average is displayed in inset above (rat 6–5, session 2409, unit 0-1); scale bars, 50 μV vertical, 1 ms horizontal.

significant, in the histograms the first peak appeared to be wider while later peaks at the same site appeared to be narrower and more clearly defined. This is consistent with the possibility that the initial peak could also be associated with the approach behavior.

Ten of the 14 neurons with anticipatory activity showed significant peaks for this 'erroneous prediction' (see supporting Fig. S2 for more examples). Of these, eight neurons were TFNs (as in Fig. 5) while the remaining six were PFNs (as shown in Fig. 4). These neurons were found with similar incidence in the accumbens core (n = 5) and shell (n = 8; P = 0.41, df = 1, χ^2 test; for one neuron the histology was unclear). While no such cells were found in the caudate, electrode placements often bypassed this area as ventral striatum was the target of this study.


Reward delivery
55Hz
*
*
-2
-1
0
1
2
3
Time (sec)
Reward delivery
55Hz
* * * * * *
-2
-1
0
1
2
3
4
5
6
7
Time (sec)
Reward delivery
55Hz
* * * *
-2
-1
0
1
2
3
4
5
Time (sec)
Reward delivery
55Hz
* *
-2
-1
0
1
2
3
4
5
6
7
8
9
Time (sec)
Reward delivery
55Hz
-1
0
1
Time (sec)

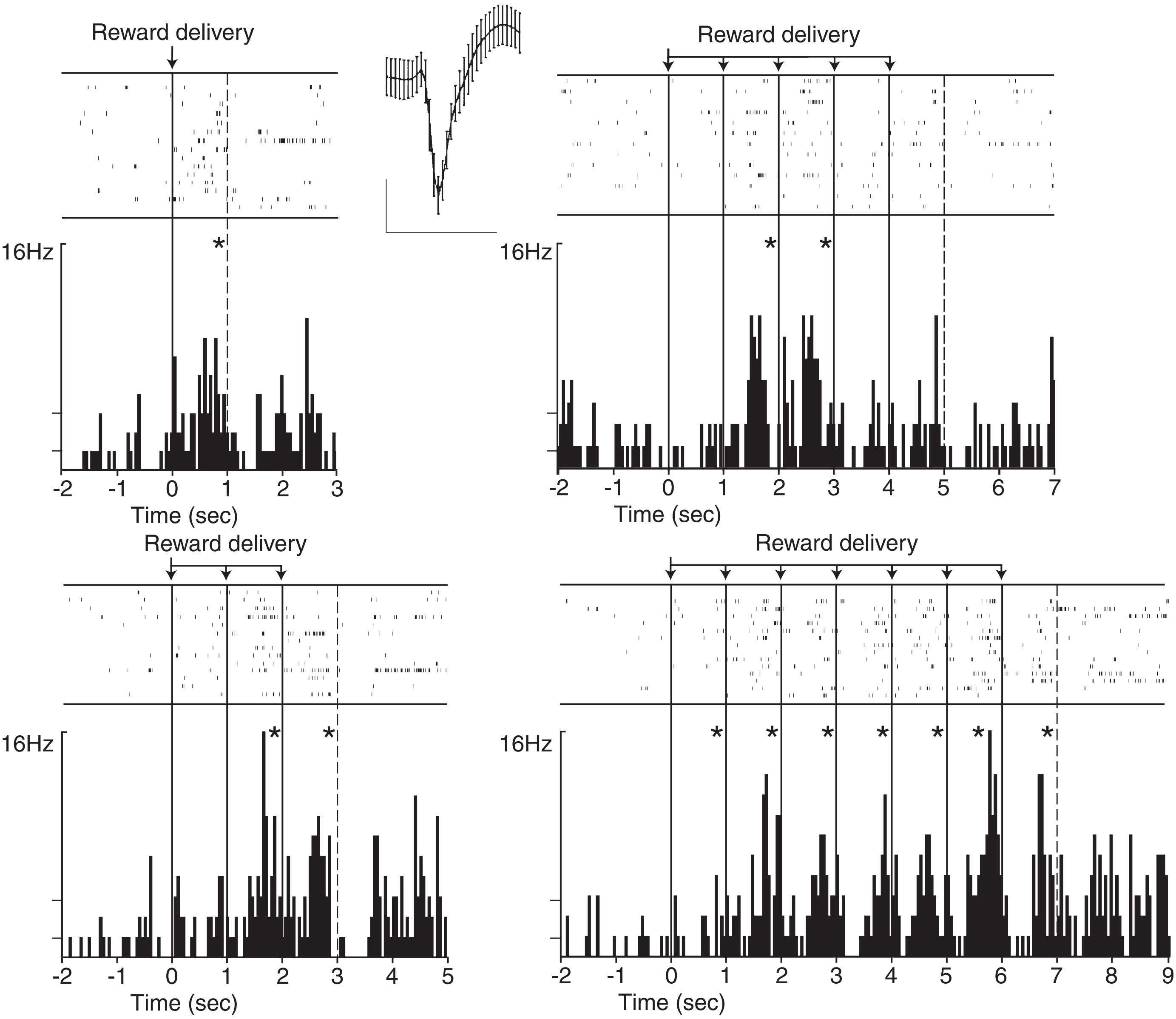


FIG. 6. Phasic excitatory activity anticipating reward delivery in a nucleus accumbens core neuron. This neuron discharged little for the first two droplets of water. The activity was greatest for the final droplet of water and for the corresponding period 1 s after the final droplet (corresponding to inappropriate anticipation of another reward). This neuron was distinguished from others in this group by a rather low firing rate. Discharges started during the 800 ms preceding water rewards. Format is same as that in Fig. 4. Waveform average is shown in inset above (rat 6-5, session 1709, unit 0-1). Scales bars, 50 μV vertical, 1 ms horizontal.

Neurons in this group had particular preferential selectivities for the order of presentation of water droplets: early, in the middle or late in the sequence. However, none of these neurons showed the decremental activity predicted by the models of Montague *et al.* (1996) and Suri & Schultz (2001). Figure 6 is an example of a PFN that had only minor responses anticipating the first and second droplets of water (one-way ANOVA, $F_{2,81} = 2.49$, $P > 0.05$), but significant activity for the final of multiple rewards (one-way ANOVA, $F_{1,59} = 9.0$, $P < 0.01$). Another pattern appeared in two neurons which fired maximally prior to and during delivery of the fourth droplet of water (cells 631509 and 631609 in supporting Fig. S2) exceeding the responses for the first or last droplets (one-way ANOVAs, $F_{3,92} = 5.33$, $P < 0.01$ and $F_{3,100} = 2.76$, $P < 0.05$ respectively; *post hoc* paired comparisons, $P < 0.05$).This variability demonstrates an uncoupling between presumed level of behavioral anticipation or expectation and the activity of individual neurons. The first drop of water should have been anticipated with a very high degree of certainty, yet there is little such anticipatory activity in the neuron of Fig. 6. Yet other neurons had

FIG. 5. A tonically active ventromedial caudate neuron with phasic excitatory activity predicting and following rewards. The response was greatest for the first droplet of water and lower for the final droplets, but the weakest responses were found for intermediate droplets. Moderately high activity also appeared 1 s after the final droplet was delivered. The neuron discharged from ~200 ms prior to reward trigger until 300 ms afterwards. Only data from the first half of the session are shown here; the remaining data show similar properties. (Lower panel) Data from all reward sites for the entire session are displayed at an expanded time scale to demonstrate the fine structure of the activity during delivery of the first droplet of water. Four peaks appear centered on −180, 0, 200 and 520 ms relative to the instant the first droplet of water was released. In contrast, such fine structure was not discernible for later droplets of water (in the upper panels where the activity is a broad peak centered about the water delivery). Format is same as that in Fig. 4. Waveform average is shown in inset at top (rat 6-2, session 1609, unit 2-2); scale bars, 50 μV vertical, 1 ms horizontal.

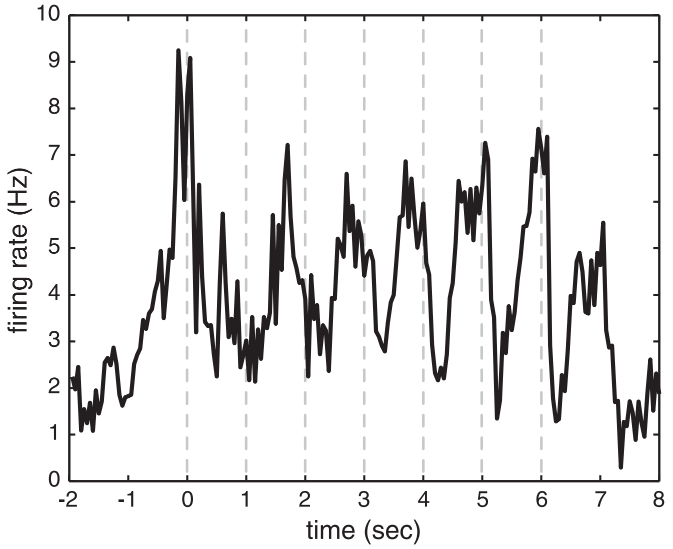


FIG. 7. Averaged activity of the fourteen ventral striatal reward-predicting neurons recorded (bins of 50 ms). The resulting trace shows characteristics similar to the pattern of individual neurons: a peak with a smaller amplitude in anticipation of a reward after the final one, and the absence of a decrement in response amplitude during the reward delivery period. The trace shows a peak with a smaller amplitude before the second droplet reward.

different order preferences while others showed no demonstrable preference for any of the droplets (cells 542307 and 652109 in supporting Fig. S2; one-way ANOVAs, $F_{7,253} = 0.35$, $P > 0.05$ and $F_{7,160} = 1.20$, $P > 0.05$ respectively). Figure 7 shows the averaged activity of the fourteen ventral striatal reward-anticipatory neurons reported here. Interestingly, in the averaged signal the peak for the second droplet reward had a smaller amplitude than those for previous and succeeding droplets, and also a peak for the erroneous reward prediction (*post hoc* paired comparisons, $P < 0.05$). The following section presents the other type of reward-related activities recorded in the ventral striatum in this task, then we will adapt a TD learning model to be compatible with these observations in a biologically plausible manner.

Activity increase during release of only the first droplet of water

In the neuron of Fig. 8 (top), the firing rate started to increase 100 ms prior to when the rat blocked the photodetector at the reward site and the activity peaked at ~150 ms afterwards. No further activity was observed for the following droplets of water at the same site (data are shown for all trials of the session). While neurons in this group varied in the onset time (from 1 s prior to arrival until slightly after arrival) and the offset time, the activity was only observed for the first droplet of water. These neurons thus would not provide a reliable signal for reward anticipation. However, in the case of single rewards this type of response would probably be confounded with those in the previous section. Interestingly, a higher proportion of such neurons were found in the accumbens core ($n = 12$) than in the medial shell ($n = 3$) and ventromedial caudate ($n = 4$; $P < 0.05$, df = 1, χ^2 test).

Uniform increase or decrease in firing rate while multiple droplets of water were delivered.

Figure 9 is taken from a TFN with inhibited activity while the rat consumed rewards. This response profile strikingly resembles tonically active neurons (TANs) reported in the monkey striatum (see *e.g.*, Apicella *et al.*, 1996). In contrast with neurons of the first group presented above, here inhibition persisted during only 1 s after the final droplet was delivered and did not continue for an additional 'erroneous' second. While this suggests that the response is correlated with the actual presence of reward, it must be noted that the onset of the inhibition began the instant the reward delivery was triggered, immediately prior to when the water would have entered the rat's mouth.

The autocorrelation analysis of this neuron's activity at the bottom of Fig. 9 demonstrates a strikingly regular timing. Note that the principal peak occurred at 0.2 s, corresponding to a frequency of 5 Hz. Other TFNs with inhibition during reward had irregularly timed and bursty activity as shown in Fig. 5. All neurons in this third group were TFNs. Most of these cells were found in the medial shell ($n = 7$), while only one such cell was found in the core and in the ventromedial caudate (χ^2 test, df = 1, $P < 0.05$).

Reproducing these reward-anticipatory responses with TD learning

This study aimed to determine whether reward-predictive activity in the rat ventral striatum is compatible with the role of this structure in TD-learning models (Sutton & Barto, 1998; see Appendix) wherein reward signals reinforce neural circuits mediating action selection. The TD learning algorithm is usually implemented within an Actor–Critic architecture, where the Actor represents a neural network that learns to select actions, while the Critic is a network that learns to compute predictions of reward. The latter reward predictions are then compared to actual rewards so that appropriate actions are reinforced, and so that the Critic's reward predictions become more accurate.

While existing models were effective for cases of single rewards, multiple rewards are more challenging as, in an ensemble of models, the striatal reward-prediction signal drops to zero the instant the first reward arrives (Barto, 1995; Foster *et al.*, 2000; Baldassarre, 2003). The few models that were tested with a temporally prolonged, but single, reward (Montague *et al.*, 1996; Suri & Schultz, 2001) hold that reward-prediction signals should progressively decrease while the animal consumes successive rewards and then finally disappear at the final reward (similar to the black trace in Fig. 10A). Whereas some cells in the monkey striatum have been found with such a decrease in activity during reward delivery (Suri & Schultz, 2001), none of the rat ventral striatal neurons that we recorded had this response pattern.

In order to replicate the absence of diminishing responses to successive rewards, the present modelling work uses the same TD-learning model while changing the model's input information. Here there is no access to the 'complete serial compound stimulus', and hence the Actor–Critic model cannot discriminate between the consecutive states that precede each successive reward. Moreover, in order to reproduce the erroneously predictive activity and variations in responses to the successive rewards, we adopt here a model composed of multiple modules of Actor–Critics which have different levels of access to precise visual and spatial input information. This is consistent with previous approaches employing multiple-module reinforcement learning models where each neuron does not encode the whole reward value function by itself (Doya *et al.*, 2002; Baldassarre, 2002; Khamassi *et al.*, 2006). This seemed to be a biologically plausible solution as it depends on the intuitive assumption that all striatal neurons do not have complete access to all spatial, temporal and visual information. This is supported by neuroanatomical studies showing inhomogeneities of terminals of corticostriatal projections, albeit within topographical delimited zones (Selemon &

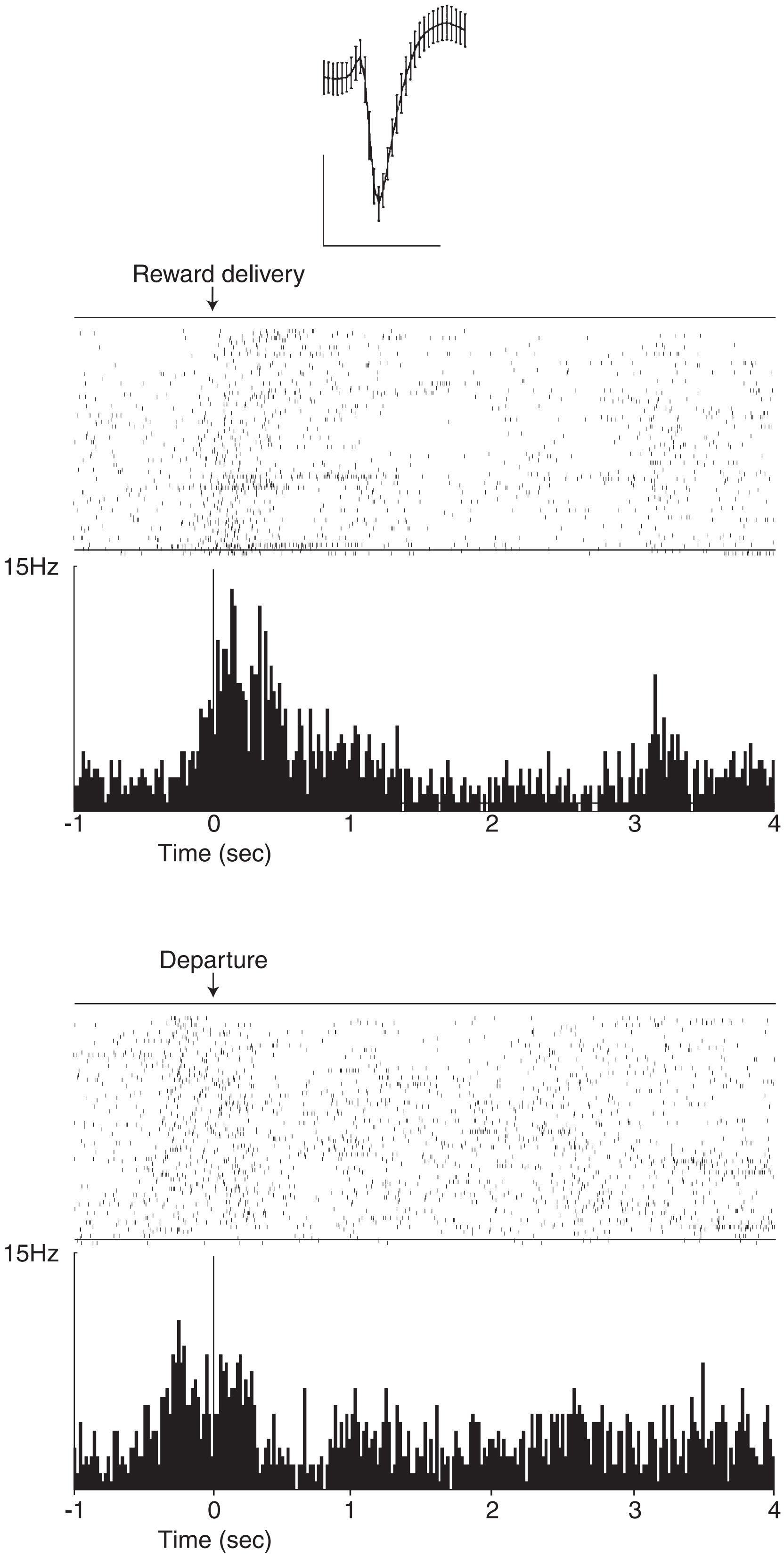


FIG. 8. This ventromedial caudate neuron discharged prior to and after delivery of the first droplet of water, but had no response to any other successive droplets. Data are shown for the whole session – only the first reward is indicated. At least two peaks are discernible here, centered on 100 and 350 ms following reward delivery. This neuron was exceptional in that it showed an increase in firing rate to ~8 Hz prior to and during departures from the water troughs (shown below). Average waveform is shown in inset at the top (rat 6-2, session 0809, unit 1-1). Bin width 20 ms; scale bars, 50 μV vertical, 1 ms horizontal.

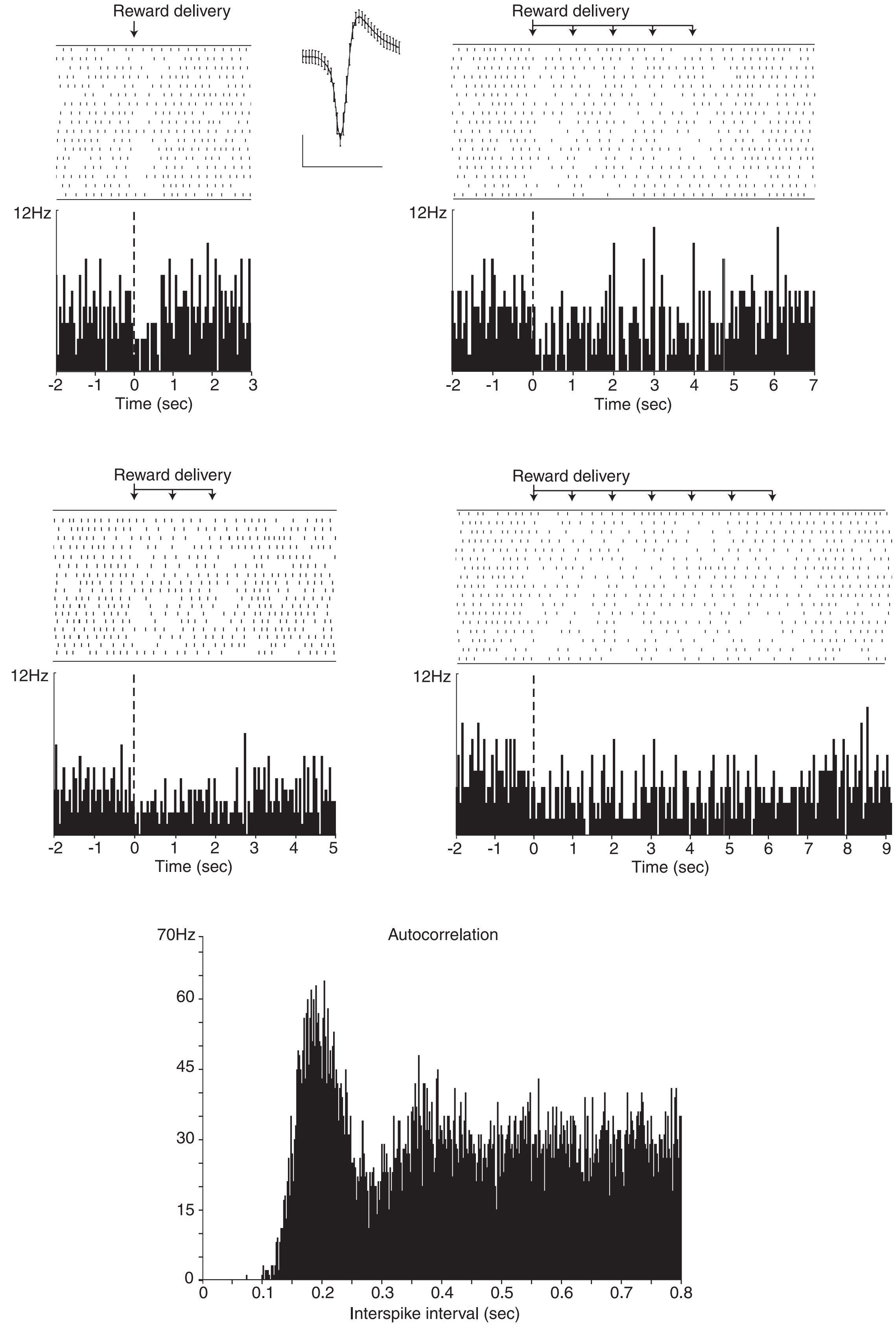


FIG. 9. Inhibition during rewards in a tonically firing neuron. The tonic activity at ~8 Hz diminished to ~3 Hz while the rat was at the reward trough consuming and waiting for more water droplets. Unlike the neurons described above, the activity resumed during the second following the final droplet of water and there was no prolongation for an additional second. Below, the interspike interval histogram has a peak at 0.2 s, corresponding to regular firing at 5 Hz. Waveform average appears in inset at the top (rat 6-1, session 1209, unit 0-1); scale bars, 50 μV vertical, 1 ms horizontal.

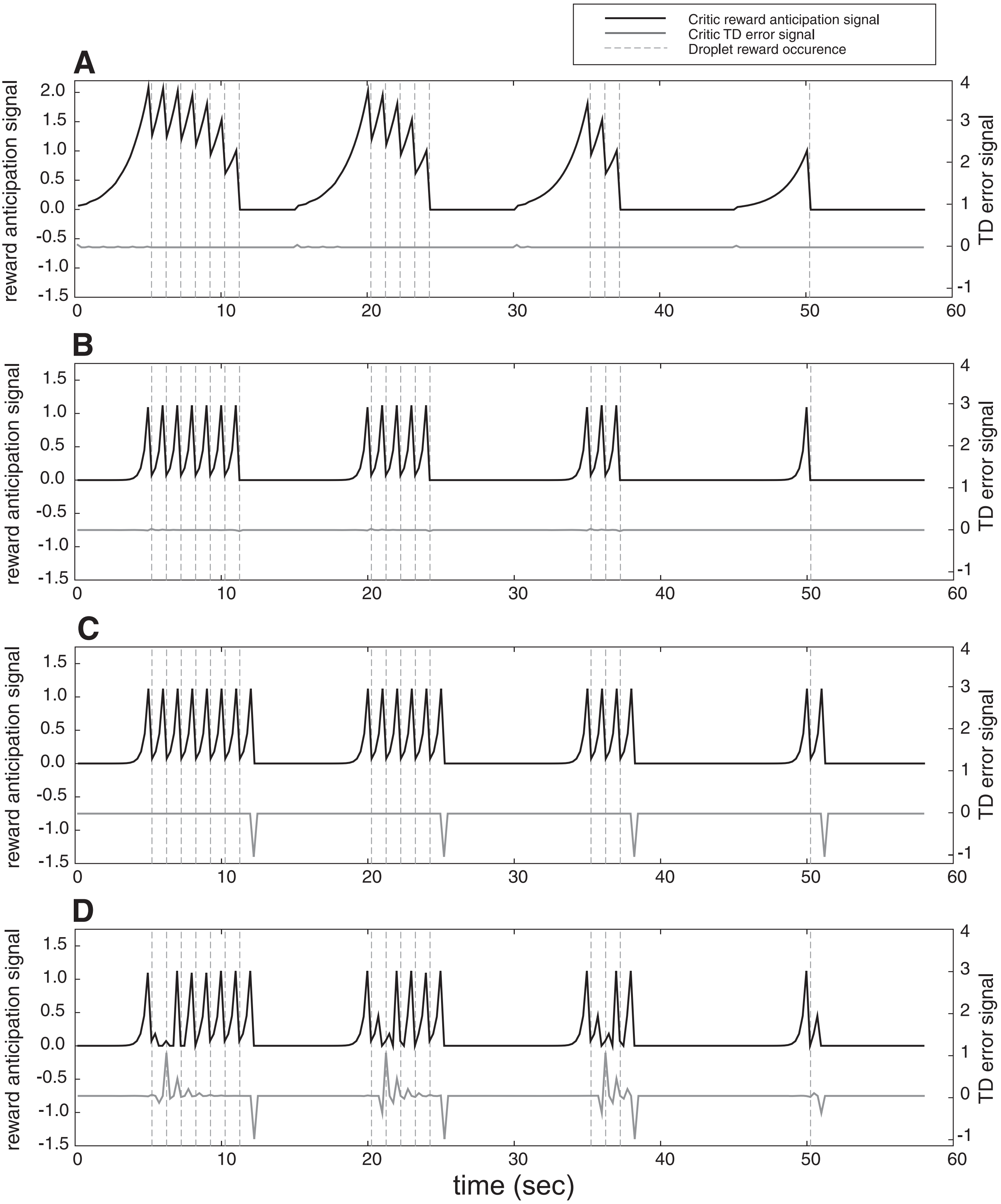


FIG. 10. Simulations of cell activity in four different versions of the TD learning model with varied inputs concerning state (spatial and sensory information), temporal-order inputs and discount factor (related to how far in the future predictions are made). The ordinate indicates average firing rate and the abscissa is time. The vertical dashed gray lines indicate the onset of rewards and the displays show successive visits to reward sites on the four arms in order of descending reward volume. The black traces show the reward-prediction signal produced by each version of the model after 25 trials of training. The gray traces show the post-training reward-prediction error signal (i.e. the TD error) associated with each reward prediction. (A) These parameters permit the model to replicate the results of Suri and Schultz (2001). (B–D) Reducing the discount factor and changing state and temporal-order inputs reproduces several of the activation patterns recorded in ventral striatal neurons. In simulations A and B the post-training TD error signal is null, consistent with recordings of dopaminergic neurons in monkeys (Schultz *et al*., 1997). In contrast, simulations C and D show non-null TD errors due to reward predictions based on incomplete task-related input information.

Goldman-Rakic, 1985; Groenewegen *et al.*, 1987; Pennartz *et al.*, 1994). Hence, neurons could belong to different 'modules' dedicated to learning part of the reward value based on different input information (Chavarriaga *et al.*, 2005; Uchibe & Doya, 2005).

The present TD-learning model has three Actor–Critic modules (it is possible to have more, but this will be sufficient to resolve the present issues). Each module independently processes the same TD-learning algorithm based upon a different mix of spatial and visual inputs. The spatial information here, that is, the state S of the animal, consists of its position (i.e., location of the respective maze arms relative to one another and the room) and sensory cues such as cue lights at reward sites. All model variants had full access to signals concerning position along the path between the maze center and the ends of the arms.

We simulated each module on 25 trials where the rats visited each of the four maze arms (actual training of the rats required many more trials). For each trial, the TD-learning algorithm was computed once every 250 ms and was simulated over the four reward sites successively. For each water reservoir, the simulation started 5 s before the first droplet reward and ended 2 s after the last. We do not address the issue of how the Actor part of the model helps to build appropriate behavior for task resolution, as this was done in a previous robotics simulation (Khamassi *et al.*, 2006). Note that the ventral striatum projects to the VTA (Thierry *et al.*, 2000) where these signals would presumably influence dopaminergic neurons, an output process out of the scope of the present modelling work. However, we will keep track of temporal-difference errors produced in the model as they constitute an experimental prediction of how dopaminergic neurons should respond in this task. Nevertheless, the goal here is mainly to study whether and how the Critic could learn to anticipate rewards in a manner similar to ventral striatal neurons, in conditions like those confronted by the rats in our task: facing the reservoir and waiting for successive rewards while a light stimulus was maintained on until the last droplet of water. As this happened only during correct trials in the real experiment (error trials were aborted to encourage the trial-and-error learning), in the simulations, the Actor part of the model had a fixed repetitive behavior.

For comparison, we reproduce the prediction of Suri & Schultz (2001) model in our task by adding the 'complete serial compound stimulus' component (Montague *et al.*, 1996). Similarly to Suri & Schultz's (2001), the discount factor γ, which indicates the capacity to take future rewards into account (cf. Appendix), was set to a high value (close to 1; here 0.85) so that the model can predict reward several seconds in the future. Figure 10A shows the resulting trace of this simulation. The long lead in initial reward-predictive activity reflects the elevated discount factor, and the gradual decrement of response strength results from the accuracy of the temporal-order signal provided by the 'complete serial compound stimulus'. In contrast, our model's first module (Fig. 10B) has no temporal-order inputs but has precise state information. This results in a series of activity peaks with identical amplitudes before each reward, and no prediction of a reward after the final one. The discount factor is set to 0.40 in order to better resemble the neural activity reported here, which commonly started ~200–800 ms before the first reward droplet. (The process of resetting discount factors could be a result of learning in the cortical–striatal–tegmental loops and is not dealt with in the model.) This results in an initial onset of predictive activity that starts later than in the previous model. The activity profile produced by this module does not have any 'erroneously predictive' activity increase after the final water droplet (similar to cell 651509-21 in supporting Fig. S2; Pearson correlation test, $n = 112$, $r^2 = 0.50$, $P < 0.001$). This type of activity could be reproduced by instantiating a second module that processes highly ambiguous state information which renders it incapable of discriminating the consecutive states following each reward droplet. As a consequence the state does not change when the light goes off in the reward alcove at the end of reward delivery while the rat remains immobile. Figure 10C shows the erroneous reward-prediction signal produced by this module (which also maintains the discount factor at 0.40). Interestingly, in some recordings, erroneously predictive activity occurred on only a fraction of trials, suggesting that such processing could be subject to gating or other modulation. The model's erroneous reward-prediction results in a non-null TD error (grey trace in Fig. 10C). This contrasts with simulations shown in Fig. 10A and B where the TD error remains null during the whole reward consumption period. Finally, Fig. 10D demonstrates how variations in selective activity among successive droplets can appear by varying spatial inputs. In this module, there is still ambiguous state information (and thus imprecise spatial information), and again the discount factor $\gamma = 0.40$. This module cannot discriminate in which of the four maze reward boxes the rat is. As a consequence, during the simulation of the model, when experiencing the water reservoir with only a single droplet reward (at time 50 s on Fig. 10D), the model learns to predict a single reward. Then, at the next trial, as the task presents successive rewards in decreasing volumes, the model experiences the water reservoir with seven droplet rewards (time 5 s in Fig. 10D). The model confounds the maze boxes (due to the absence of spatial information in the model) and predicts only a single reward. This results in an absence of peak anticipating the second droplet (time 6 s in Fig. 10D) and results in a prediction error signal at the level of 'dopaminergic neurons' in the model (grey line in Fig. 10D). We observed some ventral striatal cells with a profile of activity similar to this third module (e.g. in supporting Fig. S2 cell 523107-01, Pearson correlation test, $n = 112$, $r^2 = 0.59$, $P < 0.001$; and cell 631509-11, Pearson correlation test, $n = 112$, $r^2 = 0.59$, $P < 0.0001$). Thus these modelling results suggest that the fluctuation in the peaks' amplitude at middle droplets is due to an absence of information in these cells concerning the animal's position (this could be confirmed by analyzing whether such cells' activity is locked with hippocampal activity). Interestingly, the reward-prediction signal produced by the third module also strongly resembles the profile of the estimated population activity derived from the average over the 14 reward-predicting cells (Fig. 7; Pearson correlation test, $n = 40$, $r^2 = 0.7$, $P < 0.000001$).

Discussion

Here striatal neurons were recorded in rats as they received multiple rewards at 1-s intervals on the respective arms of a plus-maze. The experimental design aimed to disambiguate activity associated with reward-directed behaviors from actual anticipatory activity predicted by Actor–Critic models of TD learning. We found the latter in the form of phasic increases in firing rate anticipating and accompanying delivery of individual droplets of water, a novel finding in the rat striatum. This contrasted with other responses more probably associated with reward site approach behaviors and associated sensations, which took the form of phasic increases (sometimes coupled with decreases) in firing rate for the first droplet of water only.

The anticipatory lag varied among individual neurons, commencing from 800 up to 200 ms prior to the reward. Previous studies have generally shown accumbens responses that begin immediately after reward delivery (Lavoie & Mizumori, 1994; Martin & Ono, 2000; Wilson & Bowman, 2004), but in some cases precede rewards by 300 to 500 ms (Nicola *et al.*, 2004; Taha & Fields, 2005), and even as much as 1–2 s (Schultz *et al.*, 1992; Tremblay *et al.*, 1998; Shibata *et al.*, 2001; Janak *et al.*, 2004). However, as only single rewards were provided in those experiments it is not clear whether this activity in rats might be associated with sensory cues or behaviors preceding reward acquisition, or rather are actually associated with reward-anticipation only. In the immobile awake monkey preparation (Cromwell & Schultz, 2003) and in humans (O'Doherty *et al.*, 2004), however, it has been easier to reduce the risk of such confounds.

The regular timing of these anticipatory reward responses in the absence of any explicit trigger stimulus suggests that these neurons have access to some kind of timing signals (that can be reflected by the discount factor in the model). One possible source for this would be TFNs such as the one shown in Fig. 9. The highly regular 5-Hz discharges could provide a reliable basis for such timing. In this neuron the interspike interval histogram had a peak at 0.23 ± 0.10 s during the animal's running period, and this broadened and shifted to 0.29 ± 0.14 s during the reward consumption period. Interestingly, three of the peaks observed for the first droplet of water in the neuron of Fig. 5 (bottom) also had 200-ms intervals (5 Hz) between them.

Implications for models of reinforcement learning

The present results bear on recent theories and models of mechanisms of goal-directed learning engaging basal ganglia activity (Schultz *et al.*, 1997; Graybiel, 1998). The TD-learning algorithm (Sutton & Barto, 1998) has been successfully employed in neuromimetic Actor–Critic architectures to endow robots with reinforcement learning capacities (see Khamassi *et al.*, 2005 for a review). In the original formulation, the striatum makes successive predictions of reward, whose accuracy is used to compute an error prediction signal at the level of striatal-afferent dopaminergic neurons (Houk *et al.*, 1995). This prediction error, combined with signals of the presence or absence of reward, would then enable dopaminergic neurons to emit reinforcement signals that in turn modify corticostriatal synaptic plasticity. Such modifications would lead to learning by increasing the probability of selecting an action that previously led to a reward. Modification of behavior following TD-learning rules has already been observed in rats (see Daw *et al.*, 2005 for a review) and monkeys (Samejima *et al.*, 2005) during reward-based habit-learning tasks, and reward-anticipatory activity has been reported in the striatum (Kawagoe *et al.*, 1998; Miyazaki *et al.*, 1998; Daw *et al.*, 2002; Cromwell & Schultz, 2003; Setlow *et al.*, 2003). The present results extend this by showing that the rat's ventral striatal activity reflects inaccurate temporal-order information, whereas classical TD-learning models predicted that reward-prediction signals should be anchored on a precise estimation of timebins order between stimuli and rewards in order to explain dopaminergic neurons' activity (Montague *et al.*, 1996; Schultz *et al.*, 1997; Suri & Schultz, 2001). The most dramatic consequence of the present model is that it avoids making such expensive computations. Interestingly, it has been shown recently that multiple-module TD-learning models without precise temporal-order information concerning the task and with limited afferent sensory processing could still enable learning of appropriate behaviors in a simple food-searching task (Baldassarre & Parisi, 2000) and in a plus-maze task equivalent to this study (Khamassi *et al.*, 2005, 2006). This approach demonstrates the utility of models which take into account the assertion that each neuron does not have complete access to all spatial, temporal and visual information. Although the resulting multiple-module view of striatal function is structurally more complex, it is in harmony with neuroanatomical and neurophysiological data.

The present work also confirms that the diverse striatal phasic responses anticipating multiple consecutive rewards are coherent with TD learning. Striatal responses 'erroneously' predicting another droplet of water after the last one can be accounted for in the simulations as reflecting different mixes of spatial and visual information. As a consequence, ventral striatal activity is consistent with parallel reinforcement learning systems processing varying input signals (Chavarriaga *et al.*, 2005; Uchibe & Doya, 2005). The notion that different neurons receive different mixes of input information of varying levels of accuracy is consistent with known patterns of input projections to the ventral striatum (McGeorge & Faull, 1989; Groenewegen *et al.*, 1996; Mulder *et al.*, 1998).

Different striatal neurons might also process reward information at different time scales. Additional modules could vary in the value of the discount factor γ, which indicates the capacity to take future rewards into account (cf. Appendix). This reflects the observation of the variation in lag in the predictive activity from 200 to more then 800 ms and could correspond to local differences in afferent projections or local circuitry. The present discount factor of 0.4 was selected to reproduce the mean lag observed. Interestingly, in a recent brain imaging study of humans performing a reward-motivated task, different striatal subregions were selectively active according to the discount factor that best modelled the subjects' strategy concerning short- or long-term gain (Tanaka *et al.*, 2004).

Integration of multiple modules' activities

It is still an open question as to how (and where in the brain) the activity of different reward-anticipatory neurons is integrated and combined into a single reward prediction that can be used to trigger dopaminergic neurons' unitary reward-prediction error signal, i.e. the TD-error signal (Schultz *et al.*, 1997).

As different simulated modules reproduced the activity of different VS neurons, it is possible that each module could be considered to represent distinct groups of cells within the striatum. As a consequence, components of reward prediction would remain segregated within the striatum while the integration could be computed thanks to converging projections from the striatum to dopaminergic neurons (Joel & Weiner, 2000). Alternatively, as our third Actor–Critic module could represent the reward-prediction signal averaged over the ensemble of ventral striatal anticipatory neurons reported here, it could be possible that individual neurons' predictive activity is integrated locally, within the striatum, based on collateral connectivity (Gerfen & Wilson, 1996). In both interpretations, our modeling results confirm that the anticipatory activity for multiple successive rewards is consistent with participation of the rat ventral striatum in the function of a Critic that influences dopaminergic neurons' reward-prediction error signals by means of a TD learning algorithm (Joel *et al.*, 2002; O'Doherty *et al.*, 2004).

Moreover, whereas a unitary TD-error signal is commonly used in models to train an Actor to build task-relevant behaviors, there is also a specific TD-error signal that trains each module (Doya *et al.*, 2002; Baldassarre, 2003). Our model shows the TD-error signals that are expected to be associated with each module, and these would train reward prediction in specific subgroups of VS neurons. This constitutes an experimental prediction that could be tested by recording dopaminergic neurons with varying multiple successive rewards. The ventral striatal areas recorded here send projections to the SNc and to the VTA (Haber *et al.*, 2000; Thierry *et al.*, 2000; Ikemoto, 2002). Our model predicts that subgroups of brainstem dopaminergic neurons would be associated with different TD-learning modules and would, in the same plus-maze task, exhibit differential TD-error signals in response to reward (see Fig. 10): some dopamine neurons' responses to reward should have disappeared after learning, as in the seminal study of Mirenowicz & Schultz (1994). Other dopamine neurons receiving inputs from the TD-learning module which erroneously anticipates an additional droplet of water should then have inhibitory responses reflecting a reward-prediction error. If such patterns of activity are found in dopaminergic neurons, our model shows that they would still be sufficient for reinforcement learning while being consistent with the consideration that dopaminergic responses to reward should be anchored on limited information concerning the task (Redgrave *et al.*, 1999b; Redgrave & Gurney, 2006).

Daw *et al.* (2005) have recently argued that anticipatory activity for motivated behavior in rats cannot be completely explained with TD-learning models. Thus their model employs a TD module to drive habitual behavior, and this competes with a higher level tree-search module dedicated to goal-directed behavior. The present work shows that a TD-learning-based mechanism is computationally sufficient to model the diverse anticipatory responses of the neurons reported here. However, other ventral striatal neurons recorded with the present protocol (reported in Mulder *et al.*, 2004) were active from initiation to completion of goal approach behaviors. These could be embodiments of the tree-search model as they group, or 'chunk' (cf. Graybiel, 1998), behavioral sequences from departure until the arrival at the maze center, then to the next reward site. In the present work, we did not find significant differences in the number of shell and core neurons anticipating each droplet reward, as might be expected from studies demonstrating functional differences between shell and core (Pothuizen *et al.*, 2005). This is probably due to the small sample size and that many of our recording sites were near the border of these zones where fewer functional differences may be expected (Voorn *et al.*, 2004). However, the dichotomy in reward anticipation and goal approach correlates is consistent with the hypothesis that functionally distinct groups of rat nucleus accumbens neurons could be differentially involved in TD-learning or in goal-directed behavior (Dayan, 2001).

The reward-related activity observed here could serve as a Critic signal to help establish functional circuits (by a loop through VTA) for sequencing the activity of the 'goal appproach neurons' (Mulder *et al.*, 2004), first orchestrating then automating the sequence of successive steps to satisfy task exigencies. Dopaminergic neurons within this system would also transmit reward signals to more dorsal striatal regions implicated in habit learning (Haber *et al.*, 2000). Selection among alternative goal choices or even among cognitive strategies would thus be carried out in associative and limbic regions situated more ventrally in the striatum (Shibata *et al.*, 2001). This could lead to a hierarchy of behavioral control which might lead to cognitive correlates, for example, context- or reward-dependence in the more dorsal basal ganglia responses, thus participating in the translation of motivational signals into motor outputs (Mogenson *et al.*, 1980; Hikosaka *et al.*, 1989; Redgrave *et al.*, 1999a; Yin & Knowlton, 2006).

Supporting information

Additional supporting information may be found in the online version of this article:

Fig. S1. Average waveforms for each of the neurons described here.

Fig. S2. Further examples of neurons with reward anticipatory activity.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Acknowledgements

Thanks to Alain Berthoz for indispensable support throughout all stages of this project, Dr M. Zugaro for assisting with the recordings and comments on the manuscript, F. Maloumian for figure preparation, S. Doutremer for histology, S. Lemarchand for animal care and Drs K. Benchenane, J.-M. Deniau, K. Doya, A. Guillot, M.D. Humphries and A.-M. Thierry for helpful discussions. Portions of this work have been presented as posters at meetings. Grants: Human Frontiers Fellowship and the French Embassy in the Netherlands for support (A.B.M.); Cogniseine; the European Community Human Capital and Mobility program Groupement d'Intérêts Scientifiques, ACI, European Community Integrated Project ICEA-FP6-IST-27140, European Community Integrated Project BACS-FP6-IST-027819.

Abbreviations

Acb, nucleus accumbens; PFN, phasically-firing neuron; SNc, substantia nigra pars compacta; TD, temporal difference; TAN, tonically active neuron; TFN, tonically-firing neuron; VTA, ventral tegmental area.

References

Albertin, S.V., Mulder, A.B., Tabuchi, E., Zugaro, M.B. & Wiener, S.I. (2000) Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behav. Brain Res.*, **117**, 173–183.

Alexander, G.E., Crutcher, M.D. & DeLong, M.R. (1990) Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Prog. Brain Res.*, **85**, 119–146.

Apicella, P., Legallet, E. & Trouche, E. (1996) Responses of tonically discharging neurons in monkey striatum to visual stimuli presented under passive conditions and during task performance. *Neurosci. Lett.*, **203**, 147–150.

Baldassarre, G. (2002) A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviors. *J. Cog. Sys. Res.,*, **10**, 5–13.

Baldassarre, G. (2003) Forward and bidirectional planning based on reinforcement learning and neural networks in a simulated robot. In Butz, M., Sigaud, O. & Gerard, P. (Eds), *Adaptive Behavior in Anticipatory Learning Systems*. Springer Verlag, Berlin, pp. 179–200.

Baldassarre, G. & Parisi, D. (2000) Classical and instrumental conditioning: from laboratory phenomena to integrated mechanisms for adaptation. In Meyer, J.-A., Berthoz, A., Floreano, D., Roitblat, H. & Wilson, S.W. (Eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behaviour (SAB2000), Supplement Volume*. MIT Press, Cambridge, pp. 131–139.

Barto, A.G. (1995) Adaptive critics and the basal ganglia. In Houk, J.C., Davis, J.L. & Beiser, D.G. (Eds), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, pp. 215–232.

Brown, V.J. & Bowman, E.M. (1995) Discriminative cues indicating reward magnitude continue to determine reaction time of rats following lesions of the nucleus accumbens. *Eur. J. Neurosci.*, **7**, 2479–2485.

Chavarriaga, R., Strosslin, T., Sheynikhovich, D. & Gerstner, W. (2005) A computational model of parallel navigation systems in rodents. *Neuroinformatics*, **3**, 223–241.

Cromwell, H.C. & Schultz, W. (2003) Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *J. Neurophysiol.*, **89**, 2823–2838.

Daw, N.D., Touretzky, D.S. & Skaggs, W.E. (2002) Representation of reward type and action choice in ventral and dorsal striatum in the rat. *Soc. Neurosci. Abstr.*, **28**, 765.

Daw, N.D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, **8**, 1704–1711.

Dayan, P. (2001) Motivated reinforcement learning. In Dietterich, T.G., Becker, S. & Ghahramani, Z. (Eds), *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, Cambridge, pp. 11–18.

Doya, K., Samejima, K., Katagiri, K. & Kawato, M. (2002) Multiple model-based reinforcement learning. *Neural Comput.*, **14**, 1347–1369.

Foster, D.J., Morris, R.G. & Dayan, P. (2000) A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, **10**, 1–16.

Fuster, J.M. (1997) *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*, 3rd Edn. Lippincott-Raven, Philadelphia, PA.

Gerfen, C.R. & Wilson, C.J. (1996) The basal ganglia. In Swanson, L.W., Björklund, A. & Hökfeldt, T. (Eds), *Handbook of Chemical Neuroanatomy, Vol 12: Integrated Systems of the CNRS, Part III*. Elsevier Science BV, Amsterdam, pp. 371–468.

Graybiel, A.M. (1998) The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem.*, **70**, 119–136.

Graybiel, A.M. & Kimura, M. (1995) Adaptive neural networks in the basal ganglia. In Houk, J.C., Davis, J.L. & Beiser, D.G. (Eds), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, pp. 103–116.

Groenewegen, H.J., Vermeulen-Van der Zee, E., te Kortschot, A. & Witter, M.P. (1987) Organization of the projections from the subiculum to the ventral striatum in the rat. A study using anterograde transport of *Phaseolus vulgaris* leucoagglutinin. *Neuroscience*, **23**, 103–120.

Groenewegen, H.J., Wright, C.I. & Beijer, A.V. (1996) The nucleus accumbens: gateway for limbic structures to reach the motor system? *Prog. Brain Res.*, **107**, 485–511.

Haber, S.N., Fudge, J.L. & McFarland, N.R. (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J. Neurosci.*, **20**, 2369–2382.

Hikosaka, O., Sakamoto, M. & Usui, S. (1989) Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J. Neurophysiol.*, **61**, 814–832.

Hikosaka, O., Nakahara, H., Rand, M.K., Sakai, K., Lu, X., Nakamura, K., Miyachi, S. & Doya, K. (1999) Parallel neural networks for learning sequential procedures. *Trends Neurosci.*, **22**, 464–471.

Houk, J.C., Adams, J.L. & Barto, A.G. (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Houk, J.C., Davis, J.L. & Beiser, D. (Eds), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, pp. 249–270.

Ikemoto, S. (2002) Ventral striatal anatomy of locomotor activity induced by cocaine, D-amphetamine, dopamine and D1/D2 agonists. *Neuroscience*, **113**, 939–955.

Itoh, H., Nakahara, H., Hikosaka, O., Kawagoe, R., Takikawa, Y. & Aihara, K. (2003) Correlation of primate caudate neural activity and saccade parameters in reward-oriented behavior. *J. Neurophysiol.*, **89**, 1774–1783.

Janak, P.H., Chen, M.T. & Caulder, T. (2004) Dynamics of neural coding in the accumbens during extinction and reinstatement of rewarded behavior. *Behav. Brain Res.*, **154**, 125–135.

Joel, D. & Weiner, I. (2000) The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, **96**, 451–474.

Joel, D., Niv, Y. & Ruppin, E. (2002) Actor–Critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.*, **15**, 535–547.

Kawagoe, R., Takikawa, Y. & Hikosaka, O. (1998) Expectation of reward modulates cognitive signals in the basal ganglia. *Nat. Neurosci.*, **1**, 411–416.

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A. & Guillot, A. (2005) Actor–Critic models of reinforcement learning in the basal ganglia: from natural to artificial rats. *Adaptive Behav.*, **13**, 131–148.

Khamassi, M., Martinet, L.-E. & Guillot, A. (2006) Combining self-organizing maps with mixture of experts: Application to an Actor–Critic model of reinforcement learning in the basal ganglia. In Nolfi, S., Baldassare, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., Miglino, O. & Parisi, D. (Eds), *From Animals to Animats 9, Proceedings of the Ninth International Conference on Simulation of Adaptive Behavior*. Springer - Lecture Notes in Artificial Intelligence 4095, Springer, Berlin/Heidelberg, pp. 394–405.

Lavoie, A.M. & Mizumori, S.J. (1994) Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. *Brain Res.*, **638**, 157–168.

Lindman, H.R. (1974) *Analysis of Variance in Complex Experimental Designs*. W. H. Freeman and Co, San Francisco, CA.

Martin, P.D. & Ono, T. (2000) Effects of reward anticipation, reward presentation, and spatial parameters on the firing of single neurons recorded in the subiculum and nucleus accumbens of freely moving rats. *Behav. Brain Res.*, **116**, 23–38.

McGeorge, A.J. & Faull, R.L. (1989) The organization of the projection from the cerebral cortex to the striatum in the rat. *Neuroscience*, **29**, 503–537.

Mirenowicz, J. & Schultz, W. (1994) Importance of unpredictability for reward responses in primate dopamine neurons. *J. Neurophysiol.*, **72**, 1024–1027.

Miyazaki, K., Mogi, E., Araki, N. & Matsumoto, G. (1998) Reward-quality dependent anticipation in rat nucleus accumbens. *Neuroreport*, **9**, 3943–3948.

Mogenson, G.J., Jones, D.D. & Yim, C.Y. (1980) From motivation to action: functional interface between the limbic system and the motor system. *Prog. Neurobiol.*, **14**, 69–97.

Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, **16**, 1936–1947.

Mulder, A.B., Hodenpijl, M.G. & Lopes da Silva, F.H. (1998) Electrophysiology of the hippocampal and amygdaloid projections to the nucleus accumbens of the rat: convergence, segregation, and interaction of inputs. *J. Neurosci.*, **18**, 5095–5102.

Mulder, A.B., Tabuchi, E. & Wiener, S.I. (2004) Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *Eur. J. Neurosci.*, **19**, 1923–1932.

Mulder, A.B., Shibata, R., Trullier, O. & Wiener, S.I. (2005) Spatially selective reward site responses in tonically active neurons of the nucleus accumbens in behaving rats. *Exp. Brain Res.*, **163**, 32–43.

Nicola, S.M., Yun, I.A., Wakabayashi, K.T. & Fields, H.L. (2004) Cue-evoked firing of nucleus accumbens neurons encodes motivational significance during a discriminative stimulus task. *J. Neurophysiol.*, **91**, 1840–1865.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R.J. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452–454.

Paxinos, G. & Watson, C. (1998) *The Rat Brain in Stereotaxic Coordinates (CD-ROM version)*. Academic Press, New-York, NY.

Pennartz, C.M., Groenewegen, H.J. & Lopes da Silva, F.H. (1994) The nucleus accumbens as a complex of functionally distinct neuronal ensembles: an integration of behavioural, electrophysiological and anatomical data. *Prog. Neurobiol.*, **42**, 719–761.

Pothuizen, H.H., Jongen-Rêlo, A.L., Feldon, J. & Yee, B.K. (2005) Double dissociation of the effects of selective nucleus accumbens core and shell lesions on impulsive-choice behaviour and salience learning in rats. *Eur. J. Neurosci.*, **22**, 2605–2616.

Redgrave, P. & Gurney, K. (2006) The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.*, **7**, 967–975.

Redgrave, P., Prescott, T.J. & Gurney, K. (1999a) The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, **89**, 1009–1023.

Redgrave, P., Prescott, T.J. & Gurney, K. (1999b) Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.*, **22**, 146–151.

Samejima, K. & Doya, K. (2007) Multiple representations of belief states and action values in cortico-basal ganliga loops. *Ann. N.Y. Acad. Sci.*, **1104**, 213–228.

Samejima, K., Ueda, Y., Doya, K. & Kimura, M. (2005) Representation of action-specific reward values in the striatum. *Science*, **310**, 1337–1340.

Schmitzer-Torbert, N. & Redish, A.D. (2004) Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *J. Neurophysiol.*, **91**, 2259–2272.

Schultz, W., Apicella, P., Scarnati, E. & Ljungberg, T. (1992) Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.*, **12**, 4595–4610.

Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.

Selemon, L.D. & Goldman-Rakic, P.S. (1985) Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. *J. Neurosci.*, **5**, 776–794.

Setlow, B., Schoenbaum, G. & Gallagher, M. (2003) Neural encoding in ventral striatum during olfactory discrimination learning. *Neuron*, **38**, 625–636.

Shibata, R., Mulder, A.B., Trullier, O. & Wiener, S.I. (2001) Position sensitivity in phasically discharging nucleus accumbens neurons of rats alternating between tasks requiring complementary types of spatial cues. *Neuroscience*, **108**, 391–411.

Suri, R.E. & Schultz, W. (2001) Temporal difference model reproduces anticipatory neural activity. *Neural Comput.*, **13**, 841–862.

Sutton, R.S. & Barto, A.G. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge.

Tabuchi, E., Mulder, A.B. & Wiener, S.I. (2000) Position and behavioral modulation of synchronization of hippocampal and accumbens neuronal discharges in freely moving rats. *Hippocampus*, **10**, 717–728.

Tabuchi, E., Mulder, A.B. & Wiener, S.I. (2003) Reward value invariant place responses and reward site associated activity in hippocampal neurons of behaving rats. *Hippocampus*, **13**, 117–132.

Taha, S.A. & Fields, H.L. (2005) Encoding of palatability and appetitive behaviors by distinct neuronal populations in the nucleus accumbens. *J. Neurosci.*, **25**, 1193–1202.

Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y. & Yamawaki, S. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.*, **7**, 887–893.

Thierry, A.M., Gioanni, Y., Degenetais, E. & Glowinski, J. (2000) Hippocampo-prefrontal cortex pathway: anatomical and electrophysiological characteristics. *Hippocampus*, **10**, 411–419.

Tremblay, L., Hollerman, J.R. & Schultz, W. (1998) Modifications of reward expectation-related neuronal activity during learning in primate striatum. *J. Neurophysiol.*, **80**, 964–977.

Uchibe, E. & Doya, K. (2005) Reinforcement learning with multiple heterogeneous modules: a framework for developmental robot learning. In *Proceedings of the fourth International Conference on Development and Learning*, IEEE Computer Society, pp. 87–92.

Voorn, P., Vanderschuren, L.J., Groenewegen, H.J., Robbins, R.W. & Pennartz, C.M. (2004) Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.*, **27**, 468–474.

Wiener, S.I. (1993) Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *J. Neurosci.*, **13**, 3802–3817.

Wilson, D.I. & Bowman, E.M. (2004) Nucleus accumbens neurons in the rat exhibit differential activity to conditioned reinforcers and primary reinforcers within a second-order schedule of saccharin reinforcement. *Eur. J. Neurosci.*, **20**, 2777–2788.

Wilson, D.I. & Bowman, E.M. (2005) Rat nucleus accumbens neurons predominantly respond to the outcome-related properties of conditioned stimuli rather than their behavioral-switching properties. *J. Neurophysiol.*, **94**, 49–61.

Winer, B.J. (1971) *Statistical Principles in Experimental Design*, 2nd Edn. McGraw-Hill, New York, NY.

Yin, H.H. & Knowlton, B.J. (2006) The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.*, **7**, 464–476.

Appendix

The TD learning algorithm was developed in the field of optimal control theory and provides an efficient method for an embedded agent (animat, robot or other artifact) to learn to assemble a sequence of actions enabling it to optimize reinforcement (e.g. reward) in a given environment (Sutton & Barto, 1998). This approach addressed the problem that rewards may arrive considerably later than the initial neural activity, too late to modify the appropriate synapses (the ‘credit assignment problem’). TD learning has since been successfully used to describe reinforcement learning mechanisms in basal ganglia networks, but mainly for single rewards. It has been implemented in simulations where dopaminergic neurons compute reinforcement signals (Schultz *et al.*, 1997), while striatal neurons compute reward anticipation (Suri & Schultz, 2001). A given task is represented as a discretized series of timesteps. At each timestep, the agent occupies a particular position (or state) in the environment, perceives a set of signals (e.g., internal signals about motivation, or visual information about the environment), and selects an action. When the agent reaches a reward location and selects an appropriate action, it receives a reward and strengthens the neural connections leading to this state.

Instead of requiring memorization of a lengthy sequence of actions to eventually be reinforced when a reward is achieved, which is costly in terms of numbers of computations and memory requirements, the TD algorithm proposes an efficient and elegant method for reinforcing appropriate state- and signal-prompted actions towards a reward. The reinforcement signal is computed on the basis of the difference between the value of the states at two consecutive timesteps (hence the name ‘temporal-difference learning’). The value of a given state S is considered to be the value of reward which is expected (or predicted) to be received in the future, starting from this state, and is noted $V(S)$. If the action A_{t-1} is performed in state S_{t-1}, and then at time t, the expected reward value V in state S_t is higher than that of S_{t-1} [i.e. $V_t(S_t) > V_{t-1}(S_{t-1})$], then action A_{t-1} is reinforced and the value of state S_{t-1} is increased. The effective reinforcement signal that drives this learning process is given by the following equation:

$$\delta_t = r_t + \gamma V_t(S_t) - V_{t-1}(S_{t-1}) \qquad (1)$$

where r_t is the reward achieved at time t, and γ is a discount factor ($0 < \gamma < 1$) which limits the capacity to take into account rewards in the far future. At each time step t, this reinforcement signal is used to update the probability of choosing action A in state S, and to update the amount of reward that state S ‘predicts’ according to the following equations:

$$P(A_{t-1}/S_{t-1}) += \delta_t \qquad (2)$$

$$\text{and } V(S_{t-1}) += \delta_t \qquad (3)$$

where += means ‘is incremented by’.

It remains to be verified whether an algorithm of this type is actually implemented in the vertebrate brain. Nevertheless, it provides an initial intelligible framework for understanding a possible way to learn a sequence of actions towards a reward. Its simplicity and efficiency support its compatibility with the constraints of natural selection.