Title

A novel critic signal in identified midbrain dopaminergic neurons of mice training in operant tasks

Abbreviated title

VTA DA neurons predict rewards during training

Authors

Jumpei Matsumoto[1], Virginie J. Oberto [2,3], Marco N. Pompili[2], Ralitsa Todorova[2], Francesco Papaleo[4], Hisao Nishijo[1], Laurent Venance[2], Marie Vandecasteele[2] and Sidney I. Wiener[2,*]

[1]System Emotional Science, University of Toyama, 930-0194 Toyama, Japan

[2]Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, Université PSL, 75005 Paris, France

[3]Neuro-Electronics Research Flanders, 3001 Leuven, Belgium

[4]Genetics of Cognition Laboratory, Neuroscience Area, Istituto Italiano di Tecnologia, 16163 Genova, Italy

* Corresponding author: sidney.wiener@college-de-france.fr

38 pages

4 Figures, 1 Supplementary table, 4 Supplementary figures

Abstract: 248 words; Introduction: 644 words; Discussion: 1436 words

Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgements

Thanks to France Maloumian for invaluable help with figures, to Yves Dupraz for constructing the experimental chamber, Estelle Anceaume for microscopy training, Michaël Zugaro for helpful suggestions and informatics support, Drs. Mehdi Khamassi, Benoît Girard and Céline Drieu for helpful comments, Dr. Guillaume Dugué and NeuroFabLab for facilitating 3D printing

of headstages, Drs. Karim Benchenane and Liyang Xiang for help setting up the optical stimulation, and Dr. Philippe Faure for advice on recording dopaminergic neurons with chronically implanted octrodes.

Keywords

ventral tegmental area; basal ganglia; learning; temporal difference learning; actor-critic model

Abstract

In the canonical interpretation of dopaminergic neuron activity during Pavlovian conditioning, initially cell firing is triggered by unexpected rewards. Upon learning, activation instead follows the reward-predictive conditioned stimulus, and when expected rewards are withheld, firing is inhibited. However, little is known about dopaminergic neuron activity during the actual learning process in complex operant tasks. Here, we recorded optogenetically identified dopaminergic neurons of ventral tegmental area (VTA) in mice training in multiple, successive operant sensory discrimination tasks. A delay between nose-poke choices and trial outcome signals (for reward or punishment) probed for predictive activity. During training, but prior to criterion performance, firing rates signaled correct versus incorrect choices, but prior to outcome signals. Thus, the neurons predicted whether choices would be rewarded, despite the animals' subthreshold behavioral performance. Surprisingly, these neurons also fired after reward delivery, as if the rewards had been unexpected according to the canonical view, but activity was inhibited after punishment signals, as if the reward had been expected after all. These inconsistencies suggest revision of theoretical formulations of dopaminergic neuronal activity to embody multiple roles in temporal difference learning and actor-critic models. Furthermore, on training trials when these neurons predicted that a given choice was correct and would be rewarded, surprisingly, the mice adhered to other non-rewarded and untrained task strategies (e.g., spatial alternation). The DA neurons' reward prediction activity could serve as critic signals for the choices just made. This consistent with the notion that the brain must reconcile multiple Bayesian belief representations during learning.

Significance statement

The canonical view of dopaminergic function based on classical conditioning studies evokes reward-prediction error (RPE) signaling. Here, in mice performing a series of novel operant tasks with a delay between behavioral responses and reward/punishment signals, some neurons fired differentially after correct vs incorrect responses, but prior to the trial outcome (reward/punishment) signal. Nevertheless, the animals performed at chance levels, employing behavioral strategies other than the one signaled by these neurons. Furthermore, these same neurons showed canonical RPE responses, increased firing after reward signals (typically interpreted as the reward being unexpected) and firing rate decreased with

72 punishment signals (interpreted as the reward having been expected). These findings indicate
73 that dopaminergic neurons can participate in diverse functions underlying learning different
74 behavioral strategies.
75
76

Introduction

Adaptation for survival requires associating predictive cues with appropriate behavioral patterns to favor positive outcomes and to avoid aversive ones. Dopaminergic activity is implicated in the modification of neural circuits making such associations. Early in Pavlovian conditioning, dopaminergic neurons fire phasically after unexpected rewards. The DA neuron responses to rewards persist on trials immediately following learning acquisition, and then gradually diminish on subsequent trials (Hollerman & Schultz, 1998). With learning, responses appear after presentations of a conditioned stimulus (CS) that predicts the reward, and inhibitory responses occur after the expected reward is withheld.

These responses are also interpreted as evidence that dopaminergic neurons signal "reward prediction errors" (RPEs) (Houk et al., 1995; Montague et al., 1996, Schultz et al., 1997). RPE activities are posited to code "motivational value" (Schultz et al., 1997; Wise, 2005). RPE responses in dopaminergic neurons would correspond to a neural substrate for the "temporal difference (TD)" machine-learning algorithm which attributes eventual rewards or punishments to modify the circuitry leading up to these outcomes. TD learning estimates the expected reward on the basis of the current "state", that is, environmental context and cues, as well as the agent's actions. The expected reward is compared between the current state, and preceding estimates. Any differences, positive or negative, referred to as the "TD error", can then be employed to improve estimates of the state value. This "critic" dopaminergic signal would then modify activity of striatal "actors" for adaptive action selection (e.g., Khamassi, et al., 2005). These formulations originated in studies of Pavlovian conditioning, and their extension to learning adaptive behaviors, as in instrumental conditioning, still requires further elaboration regarding how appropriate actions are developed to optimize the balance of positive and aversive outcomes.

Recently, Cazettes et al. (2023) recorded in mice performing tasks with multiple possible response policies, and observed representations of unused strategies in M2 cortex, which projects to VTA (Watabe-Uchida et al., 2012). We hypothesized that if VTA neurons could then represent the rule currently being acquired during training, even while the animals are performing other strategies, since this could eventually serve as a "teacher" signal.

To test for such activity, we recorded mice as they trained in and acquired several versions of odor discrimination and visual discrimination tasks within the same experimental chamber. Trial outcome signals (signaling reward or punishment) were delayed after the

nosepoke responses. The mice were a mutant strain, permitting optogenetic identification of dopaminergic neurons with light stimulation (Eshel et al., 2015).

As anticipated, we observed phasic dopaminergic neuron activity in the 0.5-1.0 s interval after nose-poke choices but prior to trial outcome signals. Interestingly, a subset of neurons discharged in this interval distinguishing correct from incorrect choices, as observed by Schoenbaum, et al. (1998, 1999) in basolateral amygdala and orbitofrontal cortex. Thus, these responses signaled reward prediction (RP). Furthermore, this activity occurred during training, while the animals executed a variety of other strategies prior to ultimately reaching criterion performance. This activity could thus serve as a timely teacher signal. Surprisingly, most of these same neurons also had typical excitatory responses to reward delivery signals (as found prior to Pavlovian learning; Schultz, et al., 1997), and inhibition to punishment signals (as found after Pavlovian learning; Schultz, et al., 1997). Thus, the neurons signaled that the nosepoke choice was correct, and thus predicted reward, but also responded as if the rewards were unexpected, according to the canonical interpretation of dopamine's role in the RPE/TD learning framework. This may be reconciled in the perspective of instrumental learning, where the response policy could be instructed and reinforced by a dopaminergic teacher signal prior to trial outcome. These representations of alternative strategies could reflect multiple Bayesian belief representations which must be reconciled prior to reaching criterion performance levels. Functionally, dopamine release from reward prediction could prime the network for reinforcement by the subsequent RPE reward-related dopamine release, or punishment-related dopamine absence.

Materials and Methods

Animals

All experiments were performed in accordance with EU guidelines (Directive 86/609/EEC). Subjects were four transgenic adult (2-9 months) male mice expressing a channelrhodopsin2- yellow fluorescent protein fusion protein (ChR2 (H134R)-eYFP) on dopamine neurons. These mice were obtained by mating mice expressing the CRE recombinase under the control of the dopamine transporter (DAT-IRES-CRE mice, Stock 006660, Jackson Laboratory, ME, USA) with mice bearing a CRE-dependent ChR2(H134R)-eYFP gene (Ai32 mice, Stock 012569, Jackson Laboratory). Animals were housed in temperature-controlled rooms with standard 12-hour light/dark cycles and food and water were available

ad libitum. Each workday, animals were handled to habituate to human contact, and weighed. During pre-training and experimental periods, food intake, including pellets provided in the experimental apparatus, was restricted to a maximum of 3 g/day. Water access remained ad lib. Supplemental food was provided if weight fell below 85% of the normal weight.

Surgery

Anesthesia was induced with a mixture of ketamine (66 mg/kg) and xylazine (13 mg/kg) and sustained with isoflurane (0.5 – 1.0%). Mice were placed in a stereotaxic device (David Kopf Instruments) and maintained at 37° C. The scalp was exposed and cleaned with hydrogen peroxide solution. Miniature jeweler's screws were screwed into trephines, and attached with dental cement. A custom electrode assembly (developed by JM; see Oberto, Matsumoto, et al., 2023) was implanted into the left VTA and SN. Briefly, the assembly consisted of microdrives holding four recording probes for recording neural activity and two optic fibers for optogenetic identification with light stimulation. Each recording probe was twisted bundles of 8 formvar coated 12 micron diameter nichrome wires ("octrodes") were inserted in polyimide tubes. The wires were gold-plated to an impedance of about 350 kohm. The stainless-steel screws implanted on the left and right cerebellum as ground and reference electrodes, respectively.

The behavioral task

Experimental apparatus

The 'Operon' system developed by Scheggia et al (2014) was adapted as an experimental chamber permitting the mice to perform olfactory and visual discrimination tasks (Figure 1). Many components were purchased from Med Associates (Fairfax, VT, USA). Each of the short sides of the rectangular (160L x 136W x 160H mm) plexiglass chamber has two arrays of LEDs and two ports used for both odor sampling and nose pokes. An olfactometer presented d- and l- stereoisomers of limonene at the two odor ports in a pseudo-random sequence. All six white LEDs were lit above one nose-poke port and off above the other, again in pseudo-random sequence. Between the ports is a central feeder port with an overhead lamp which is lit for 7 s during the delay after error trials. This lamp emits only 55 lx, which is not intrinsically aversive. A reward pellet (5TUL TestDiet, Richmond, IN, USA, 14 mg,) was delivered by a dispenser (ENV-203-14P, Med Associates, Fairfax, VT, USA). A plexiglass

barrier divided the two sides of the chamber and was slid up from below via an automated motor assembly. Behavioral procedures followed those of the "Extradimensional shift" protocol of Scheggia and Papaleo (2016) but without textural cues, and further details can be found there and in Scheggia et al. (2014).

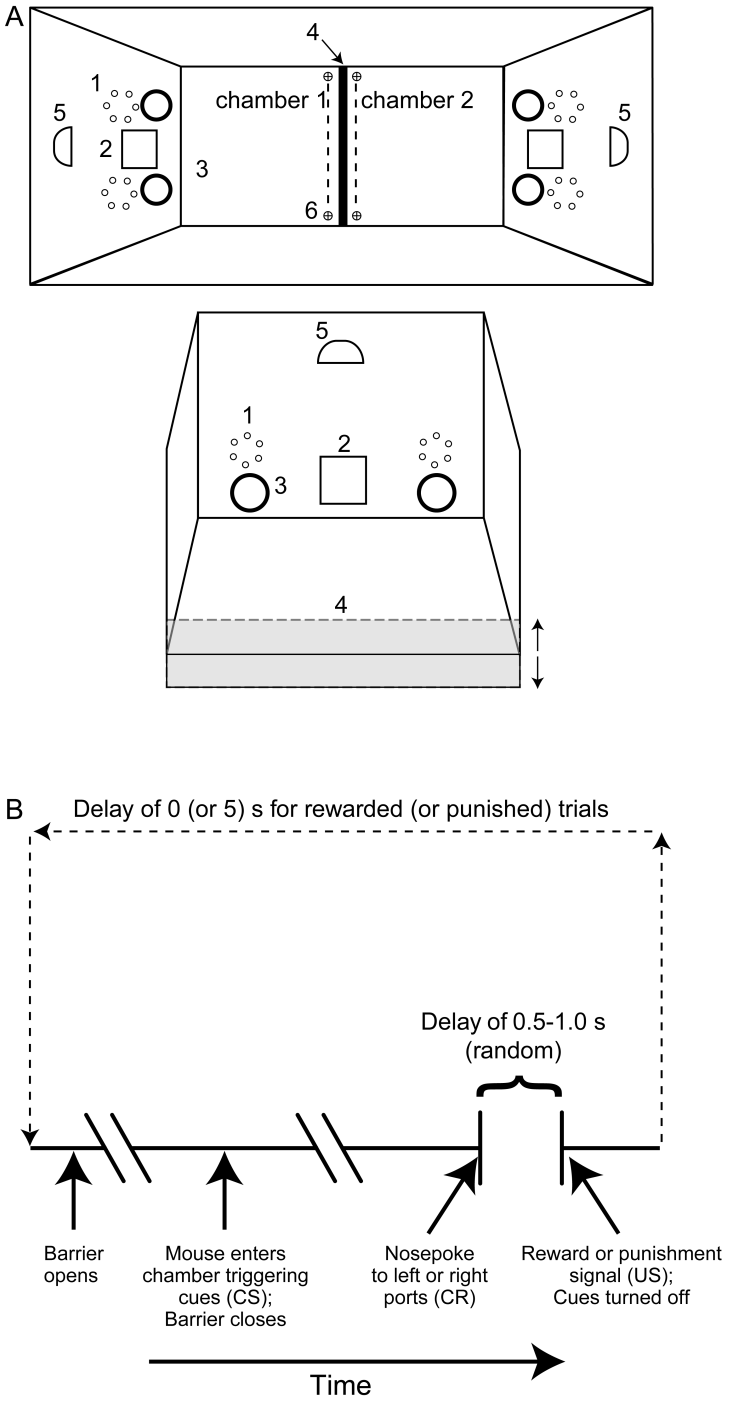


Figure 1. The behavioral apparatus and task (after Scheggia, et al., 2014).
A) Top) Overhead view. 1- Hexagonal LED array; 2- Reward dispenser; 3- Nose-poke port; 4- Moveable barrier; 5-House light; 6- Photobeam. Bottom) Frontal view of one chamber (photobeam not shown here). **B)** The automated task sequence.

Pre-training

Mice were provided water ad libitum, but mildly food deprived to maintain them above 85% their baseline weight, as controlled with daily weighing. The behavioral apparatus was turned on, and the olfactometer outputs were confirmed to present odors discriminable by the experimenter. Then, the mouse was placed into one compartment of the conditioning chamber. The room lights were dimmed and the behavioral task was started. Animals were first pre-trained to shuttle between two boxes for reward. In the first stage of training, a nose poke into either port triggered a pellet delivery. Then animals were pre-trained for simple visual discrimination task (light-on vs light off). While the rewarded side was varied pseudo-randomly, in cases of spatial persistence, the other side was rewarded more. Once they could regularly perform sequential trials in the maze, mice were provided food ad lib prior to implantation with optrode assemblies, and training was resumed along with recordings.

Behavioral task sequence

After the pre-training, animals performed the following behavioral training on each recording day. At the beginning of each trial, the barrier was lowered. Once the mouse crossed a photodetector on its way to the other half of the chamber, one of the two LED arrays was lit and the two odors were released from the respective odor ports on that side. This photodetector crossing is called the “Cue” event. Once the mouse cleared these photodetectors, the barrier was slid shut. Nose-pokes to the port on the side of the (preselected) odor or visual cue crossed a photodetector within the port that triggered food pellet release, or an error signal, at a variable latency ranging from 0.5 to 1.0 s. The position of the rewarded cue was pseudo-randomly varied on successive trials so that the same odor or visual cue were not presented more than twice at the left or the right nose-poke port. Furthermore, the same trajectory orientation across the chamber (diagonally crossing or running along a side wall) was not rewarded on more than 2 successive trials. The task control system permitted programming the cue contingencies and reward delivery, outputting time stamps. Erroneous choices triggered illumination of an overhead light during a timeout period of 5 or 8 s while the barrier remained up. Once mice attained criterion performance (8 rewarded choices out of 10 consecutive trials, or 6 consecutively rewarded trials) in the olfactory or visual discrimination tasks, the system automatically triggered a switch to the

other task. This is the same, or even more stringent, criterion than used in this type of study (e.g., Kaefer, et al., 2020; Lapiz-Bluhm, et al., 2009, Tait, et al., 2017).

As criterion performance was successive attained, the mice were cycled between the two tasks. In some sessions, it was necessary to pretrain the mice, starting with only the visual or olfactory cue discrimination before adding the other modality cues. When the mouse stopped performing trials for more than 120 s, the behavioral session of the day was ended. The weight of the rewards earned was calculated, and the required amount of chow for the daily feeding was calculated and provided.

Recording procedure

Recordings of neural signals were made with an Ampliplex system (sampling rate: 20 kHz; cutoff frequencies of the analog low-pass and high-pass filters were 0.3 and 10 kHz, respectively). After surgical implantation, electrodes were advanced as often as twice daily until neurons appeared with responses to optical stimulation. Electrodes were further advanced when discriminable units were no longer present.

In each recording day, the headstage was plugged in and the mouse was placed in the Operon system. Then, the neural signals were recorded while the mouse was performing the above behavioral task. The position of the mouse within the Operon box was detected and tracked using a LED attached on the headstage, which was captured with a video camera (30 frames/sec) installed above the system. After the mouse stopped performing the task, the headstage plug was changed, optic fibers were connected, and the animal was placed in a large plastic beaker. Using a 475 nm-laser light source (DPSSL BL473T3-100FL, Shanghai Laser and Optics Century, Shanghai, PRC), trains of 10 light pulses (pulse duration: 10 msec; light power: 10 mW max; frequency: 5 Hz) were given for 20 times with 10 sec inter train intervals under control of an Arduino-based system.

Data analyses

Spike sorting and neuron type classification

For single unit discrimination from the extracelluar signals recorded from the octrodes, offline spike sorting was carried out with KiloSort (Pachitariu et al., 2016) followed by manual curation using Klusters (L. Hazan, http://neurosuite.sourceforge.net).

To identify neurons as dopaminergic, we used the Stimulus-Associated spike Latency Test (SALT; Kvitsiani et al., 2013; Eshel et al., 2015). The test determines whether light pulses significantly changed a neuron's spike timing by comparing the distribution of first spike latencies relative to the light pulse, assessed in a 10-msec window after light-stimulation, to 10-msec epochs in the baseline period (-150 to 0 msec from the onset of light-stimulation; see Kvitsiani et al., 2013 for details). A significance level of $p<0.01$ was selected for this.

All neurons recorded from an octrode with at least one SALT+ response were considered to be in a dopaminergic nucleus. If the octrode was not advanced, neurons on the day before and the day after were also counted as in VTA or SNc, even if no SALT+ responses were recorded on those days. Similarly, if the octrode had not been advanced, and SALT+ responses had been recorded on non-consecutive days, intervening days with no SALT+ responses were still considered as in a dopaminergic nucleus. Also, when electrodes had been advanced, all recordings in days between those with SALT+ recordings were also considered to be in VTA/SNc. Neurons with SALT+ responses are considered dopaminergic. Neurons in a dopaminergic nucleus not demonstrating a significant SALT+ response are labelled SALT-, although this negative result does not conclusively show that the neuron is not dopaminergic. All neurons qualifying as in dopaminergic nuclei were categorized with clustering according to the following parameters: spike width, mean firing rate. The criterion for fast spiking interneurons (FSI) was firing rate > 15 Hz and spike width < 1.5 ms (cf., Ungless and Grace, 2012). No FSI's were SALT+. All other SALT- neurons in dopaminergic nuclei are referred to as "Other".

Behavioral correlates of neural activity

The cells with mean firing rates <1 spike/s in a given epoch were excluded from analyses. We defined that "Cue" as when the mouse first crossed the photobeam while entering the next chamber, triggering cue onset; "Choice" is the time of the nose-poke response, as detected by the photo-detectors in the ports; "Outcome" is either the instant the dispenser released the reward pellet (which made a salient sound), or the onset of the punishment period. For each neuron in each epoch, firing rate during baseline periods (0.5 s before the cue), [cue, choice] periods, [choice, outcome signal] periods, and post-outcome periods [0.5 s after the outcome signal] were calculated. The RP responsive neurons were defined as a neuron that showed significant difference(s) between rewarded and non-

277 rewarded trials in either or both of [cue, choice] and [choice, outcome signal] periods
278 (unpaired t-test, $p < 0.05$). Similarly, the responses to the reward and punishment were tested
279 by comparing firing rate in the baseline period with the outcome periods of rewarded and
280 non-rewarded trials, respectively (unpaired t-test, $p < 0.05$).
281
282 **Experimental design and statistical analyses**
283 Subjects were four transgenic adult (2-9 months) male mice expressing a
284 channelrhodopsin2- yellow fluorescent protein fusion protein (ChR2 (H134R)-eYFP) on
285 dopamine neurons. Data samples were recordings of neurons from individual periods when
286 the mice were challenged with one of several sensory discrimination tasks. (The result show
287 that a single neuron could show different response properties in the different tasks.) In order
288 to respect "Reduction" of the 3R's, experiments were stopped when sample sizes were
289 decided to be sufficient, i.e., when significant effects could be determined reliably and SEM
290 error bars were minor.
291 Basic statistical methods such as paired and unpaired t-test, chi-square test, and
292 Pearson's correlation analysis were used for statistical comparisons. The statistical tests were
293 performed using MATLAB. The significance threshold was set to $p<0.05$. The values for n, p,
294 and the specific statistical test performed for each analysis are included in the corresponding
295 figure legend, table, or main text.
296
297 **Histology**
298 Once stable recordings were no longer possible, marking lesions were made with 10 s
299 of 30 µA cathodal current. After waiting for at least 90 min, mice were then killed with a lethal
300 intraperitoneal injection of sodium pentobarbital, and perfused intra-ventricularly with
301 phosphate buffered saline solution (PBS), followed by 10% phosphate buffered formalin. The
302 brain was removed, post-fixed overnight, and placed in phosphate buffered 30% sucrose
303 solution for 2-3 days. Frozen sections were cut at 80 µm, and permeabilized in 0.2% Triton in
304 PBS for 1 h at room temperature. Antibody staining was performed to confirm localization of
305 electrodes in dopaminergic nuclei. Sections were then treated with 3% bovine serum albumen
306 (BSA) and 0.2% Triton in PBS for 1 h with gentle agitation at room temperature to block non-
307 specific binding. Sections were then rinsed for 5 min in PBS at room temperature with gentle
308 agitation. Then the sections were left overnight with gentle agitation at 4° C in a solution of

the first antibody, mouse monoclonal anti-TH MAB318 (1:500), 0.067% Triton, and 1% BSA in PBS. After three rinses for five minutes in PBS, the sections were treated with a second antibody (1/200, anti-mouse, green), Nissl-red (1:250), and 0.067% Triton in PBS for two hours at room temperature. Then sections were rinsed 3 times for five minutes in PBS, and mounted with Fluoromount®. Sections were examined with fluorescence microscopy to verify electrode tips and trajectories in the immunofluorescent stained zones (see Supp. Fig. 3D of Oberto, Matsumoto, et al., 2023). Anatomical reconstruction with fluorescence microscopy revealed RP neuron recording within the ventral tegmental area at mediolateral positions ranging from 375 to 517 µm.

For further details, see Oberto, Matsumoto, et al. (2023).

Results

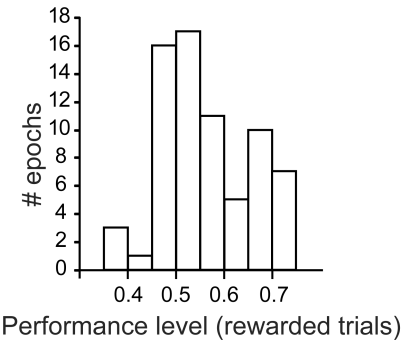
The behavioral task

Mice with optogenetically responsive dopaminergic neurons were first trained in a visual discrimination task in a double chamber apparatus adapted from Scheggia, et al. (2014) (Figure 1). At the beginning of each trial, a central barrier was lowered. Once the mouse crossed a photodetector to enter the other chamber, a cue was presented (LED lights on at the left or the right nose-poke ports in a pseudo-random sequence) for the visual discrimination task. Nose-poke choices were followed by a delay at variable latencies ranging from 0.5 to 1.0 s prior to either the reward signal (sound of a food pellet dispenser), or, alternatively, a punishment signal (for 5 or 7 seconds, chamber lights on coupled with a delay for barrier removal permitting the next trial). Once criterion performance was achieved in the visual discrimination task, mice were then implanted with multiple electrodes and equipped for intracranial optogenetic stimulation.

The mice were then recorded as they re-acquired the visual discrimination task. When criterion performance was again reached, an olfactory discrimination task was introduced in the same maze, with two different odors emanating from the nose-poke ports, again in pseudo-random sequence. In practice, when a mouse achieved criterion performance (6 successive rewarded trials, or 8 out of 10), the task was changed (see Annex of Khamassi, et al., 2024, for justification of these criteria). If only one cue modality had been presented, the other was then added without changing the reward contingency. This challenged the animal to continue performing the task in the presence of a second inconsequential cue. If criterion performance was reached while the two cue modalities were presented simultaneously, the reward contingency was shifted to the other sensory modality, and both cue types were still presented. Should the mouse reach criterion again, the reward contingency was again shifted to the other cue modality discrimination task (visual to olfactory, or olfactory to visual). Each period with one or both cue types or with different reward contingencies is referred to below as a “task epoch”, and these periods are analyzed separately since a given neuron could respond differently in different task epochs (detailed below). The reward contingency switching corresponds to the extra-dimensional attentional set-shifting task of Scheggia, et al (2014) which would require intact prefrontal cortical function (Dias, et al., 1996; Birrell & Brown, 2000; Bissonette, et al., 2008), which, in turn, depends upon dopaminergic input (see e.g., Vander Weele, et al., 2019; Di Domenico & Mapelli, 2023).

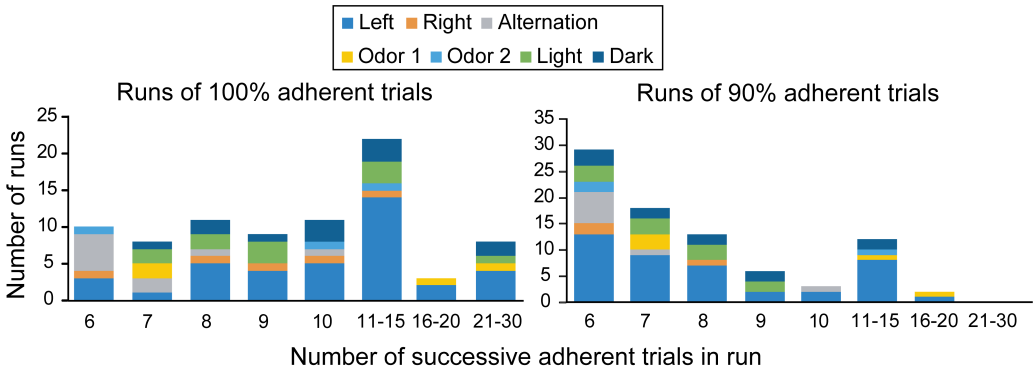
Behavior

Of the 53 task epochs when one or more reward-predictive optogenetically identified DA neurons were recorded, in 28, the mice ultimately achieved criterion performance in the task. In 18 others, the mice achieved criterion performance in another epoch in the session. Thus, learning was indeed taking place. The overall performance in these sessions was 56±1%, mean ± SEM; Supplementary Figure 1), consistent with reward contingencies being changed once criterion was reached (and thus this percentage is low since only a small fraction of trials were recorded during criterion performance).



Supplementary Figure 1. Distribution of mean performance levels (proportion of trials that were rewarded) in epochs with reward-predictive responses in identified dopaminergic neurons. N = 53 epochs. If more than one cell was recorded in a given epoch, it was only counted once.

Interestingly, as the animals were acquiring the tasks, their choices were not random. During epochs with RP responses, animals performed at criterion levels (i.e., at least 6 successive adherent trials) in strategies other than the current rule (i.e., go left, go right, spatial alternation, or choosing a currently unrewarded odor or lit/unlit nose-poke port; see Khamassi, et al., 2024, for justification of this selection of strategies analyzed). Of the 24 task epochs with RP responses and sufficient numbers of trials to analyze, all but one had runs of trials that were performed according to strategies other than the current rule (see Supplementary Figure 2).



Supplementary Figure 2. Incidences of the lengths of 247 runs of successive trials adherent to strategies other than the current rewarded rule for a sample of 24 epochs when DA reward predictive activity was recorded. (If more than one RP responding neuron was recorded within an epoch, the runs were only counted once.) These tallies included 1723 trials (See Materials and Methods for further details). Only epochs with 15 or more trials were tallied (not counting those trials qualifying for criterion performance at the currently rewarded task. For the strict "100% adherence" approach, we counted occurrences of at least successive six trials adhering to the strategy. For the more permissive "90%" adherence" approach, intermittent single non-adherent trials were allowed, as long as they were contained within a sequence of at least eight adherent trials. In cases where overlapping series adhering to two different strategies, shared trials were assigned to the longest sequence (or assigned randomly when sequences were of equal length).

Neuronal recordings

In 156 recording sessions in four mice, of total 967 neurons recorded under different task conditions, 272 were optogenetically identified as dopaminergic (DA) neurons, and 206 were fast-spiking interneurons (FSI), while 489 others could not be identified as DA or FSI. The latter are referred to as "OTHER" neurons (see Materials and Methods for criteria for category assignments). While mice were *performing at sub-criterion levels* in the sensory discrimination tasks, in the 272 task epochs with optogenetically identified DA neurons, 82 (30%) of these fired selectively according to whether the trials would be rewarded or not, but *prior to the signals indicating whether the choice would be rewarded or punished* (Figure 2). These are referred to as "reward prediction" (RP) cells. This selectivity also occurred in 106 OTHER neurons recorded in 489 epochs (21%; Supplementary Figure 3) and 31 FSI in 206 epochs (15%). The incidence of RP responses was greater in DA neurons than FSI ($\chi^2(1) = 14.8$, $p = 1.2 \times 10^{-4}$) or OTHER neurons ($\chi^2(1) = 6.74$, $p = 9.4 \times 10^{-3}$). The results reported below focus on the identified DA neurons, because of their theoretical interest. Most neurons significantly increased firing when predicting rewards than punishments (85% for DA, 55% for OTHER, and 68% for FSI; Supplementary Table 1) and only these are illustrated in the Figures and further

characterized below. In the RP cells, when the target (i.e., currently rewarded) cue was visual, 79% of significant RP responses occurred in the [cue onset, nose-poke] interval but only 31% in the [nosepoke, outcome] period in DA neurons (Figure 2, left). In OTHER neurons, 50% and 56% of significant RP responses occurred in the [cue onset, nose-poke] and the [nosepoke, outcome] periods, respectively; Supplementary Figure 3, left). In contrast, when target cues were odors, only 12% of DA (and 10% of OTHER) RP responses occurred in the [cue onset, nose-poke] while 95% (and 93% of OTHER) were in the [nosepoke, outcome] period (Figure 2, right and Supplementary Figure 3 right). In no case did the activity clearly ramp up until trial outcome signals were presented (Figure 2 and Supplementary Figure 3), in contrast with the ramping responses previously shown in dopaminergic neurons prior to rewards (e.g., Farrell et al, 2022), and which have been assimilated to a "motivational incentive" signal associated with reward seeking (Berridge and Robinson, 1998).

Supplementary Table 1. Relative incidence of RP responses in DA neurons by epoch.

DA neurons

	Visual target			Odor target			All targets		
	simple	complex	total	simple	complex	total	simple	complex	total
R>NR	16	13	29	12	29	41	28	42	70
R<NR	1	3	4	2	6	8	3	9	12
Both	0	0	0	0	0	0	0	0	0
n.s.	44	72	116	17	57	74	61	129	190
Total	61	88	149	31	92	123	92	180	272
% R>NR	26%	15%	19%	39%	32%	33%	30%	23%	26%

OTHER neurons

	Visual target			Odor target			All targets		
	simple	complex	total	simple	complex	total	simple	complex	total
R>NR	10	6	16	15	27	42	25	33	58
R<NR	11	9	20	6	19	25	17	28	45
Both	1	0	1	2	0	2	3	0	3
n.s.	71	137	208	38	137	175	109	274	383
Total	93	152	245	61	183	244	154	335	489
% R>NR	11%	4%	7%	25%	15%	17%	16%	10%	12%

R>NR (or R<NR) signifies that the average firing rate during rewarded trials was significantly higher (or lower) than during the punished (not rewarded) trials in either the [cue onset, nose-poke] or [nose-poke, outcome] period. Both means that the neuron exhibited opposite RP responses (R>NR and R<NR) in the two different periods ([cue onset, nose-poke] and [nose-poke, outcome]). n.s. – not significant.

DA and OTHER neurons with RP activity also had canonical responses to reward and punishment signals. Firing rates significantly increased after reward signals in 54/70 (77%) of DA and 40/58 (77%) of OTHER neuron epochs (unpaired t-test, $p<0.05$) while firing rates decreased significantly after punishment signals in 47/70 (67%) of DA and 20/58 (34%) of OTHER neuron epochs (unpaired t-test, $p<0.05$; columns of stars to the right of color rasters in Figure 2 and Supplementary Figure 3). Note that when this firing rate reduction occurred, it only lasted at most for 1 s after punishment onset, even though the punishment period lasted 5 to 7 s.

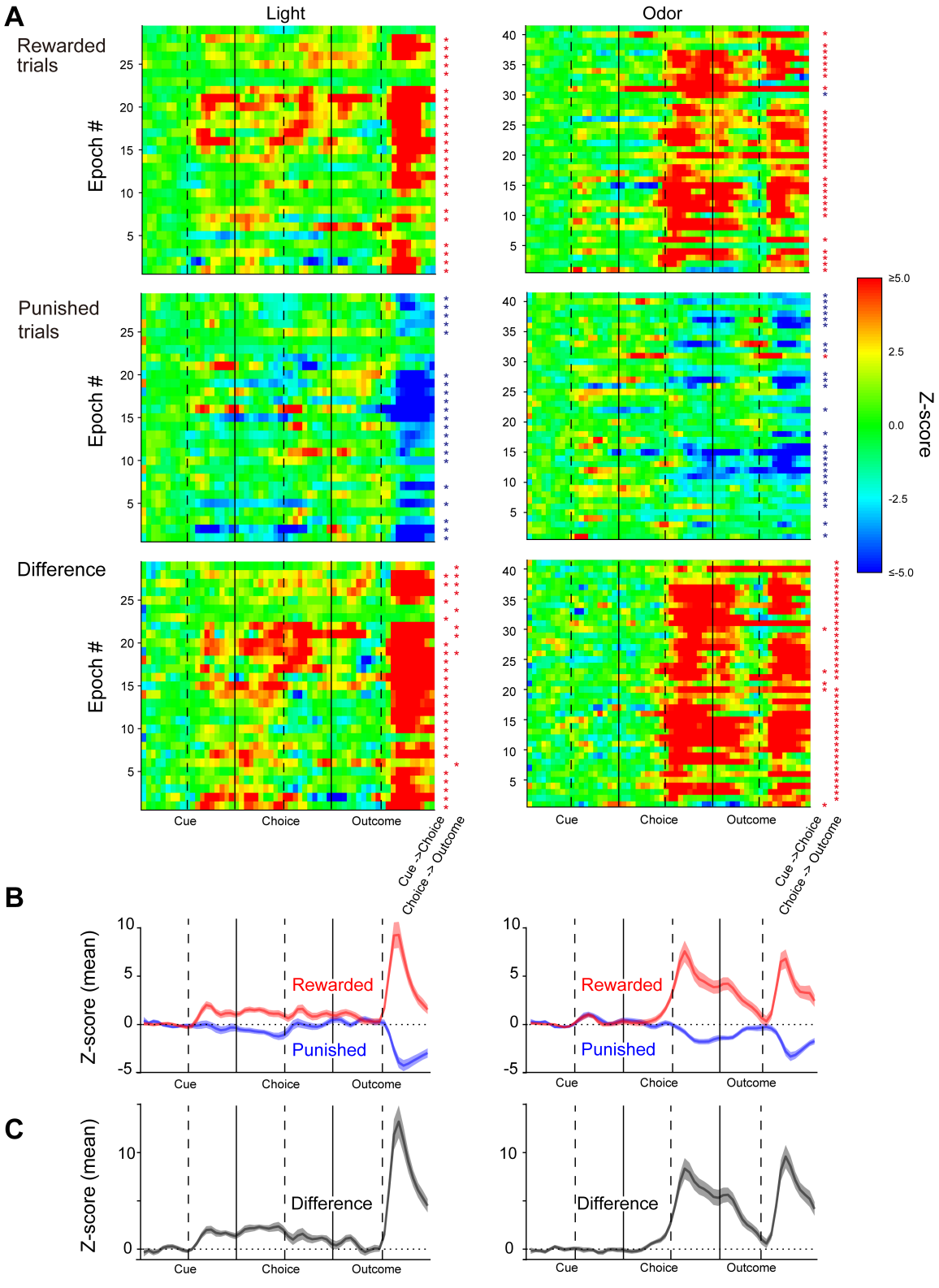
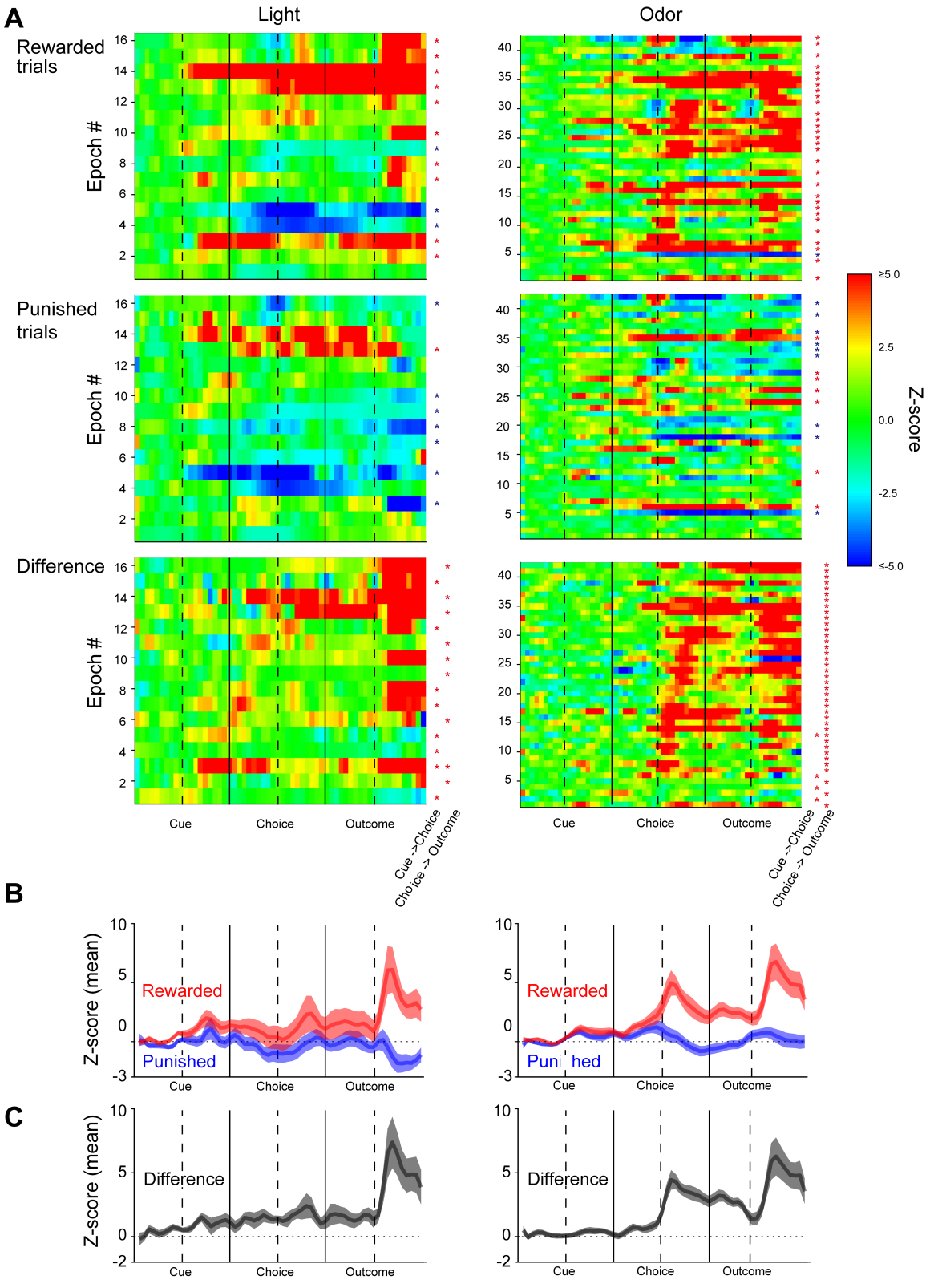


Figure 2. RP responses in DA neurons.

A, top two rows of panels) Color rasters showing z-scored firing rates in 0.5 s windows (framed by continuous vertical lines) around the three principal task events (dashed vertical lines). Each row represents firing of a neuron during a single task contingency epoch. The z-scores are calculated as bin value/[mean of all 60 bins for the cell in that epoch/SD for the 60 bins].

Right columns of stars) Canonical dopaminergic responses. In the 0.5 s periods after reward or punishment (outcome) signals, red stars indicate significant increases in firing rate relative to the 0.5 s baseline period prior to cue onset, while blue stars indicate significant firing rate decreases (unpaired two-tailed t-test, $p<0.05$). **A, bottom)** Reward prediction quantified as differences between values for rewarded and punished trials. Right) Left columns are for the [cue onset, nosepoke choice] period and right columns are for the [nosepoke choice, outcome signal] period. Here red stars indicate significant differences between firing in rewarded and punished trials (unpaired two-tailed t-test, $p<0.05$). Staircase plots are ordered based on the latency of the peak firing rate between cue and outcome in the rewarded trials in the displayed data. Note that for the statistical analyses of [cue, choice] and [choice, outcome signal] periods, all data from these periods were used, which are not necessarily represented in the figures visualizing the neural activity in 1 s windows around the events since these intervals could last longer than one second. **B)** Means of the data of the upper two rows of panels of A. Darker lines indicate mean and shaded areas show SEM. **C)** Means of differences in the third row of panels of A. N = 29 (Light) and 41 (Odor) epochs.



Supplementary Figure 3. OTHER neurons with RP responses.
Same format as Figure 2. N = 16 (Light) and 42 (Odor) epochs.

Figures 3 and 4 show examples of these RP responses in DA neurons. In addition to their prominent RP responses, firing rates significantly increased (relative to baseline) after reward onset ($p<0.05$, unpaired t-test), the canonical response previously shown in Pavlovian conditioning prior to acquisition (Schultz, et al., 1997). The cells of Figures 3B, 4A and 4B also had another canonical DA neuronal response, inhibition after the punishment onset ($p<0.05$, unpaired t-test). In the literature, this inhibition has typically been observed after the animals acquire Pavlovian conditioning, in those trials when an expected reward is withheld (Schultz, et al., 1997). Thus, in the RPE framework, this response would indicate that a reward was expected. However, the significantly lower firing rate prior to the punishment signals on these trials (relative to activity in the same period in trials to be rewarded) would indicate that no reward was expected.

In the sessions of Figures 4A and B, the mouse reached criterion twice, and, accordingly, the task was changed each time. In Figure 4A, the RP response was significant in all three of the task epochs, while the performance level averaged only 51%. But in the neuron of Figure 4B, the RP response was only significant in the third epoch (while performance was 52%). Thus, in this paper, the incidence of RP responses is tallied by task epochs rather than over whole sessions (which could dilute significant effects).

The numbers of successive trials in runs with adherence to non-rewarded strategies is shown in Supplementary Figure 2. The legends of Figures 3A, 3B, 4A and 4B describe the numerous runs of trials adhering to unrewarded strategies in the RP epochs, despite the RP responses reflecting the correct strategy. This, in conjunction with the RP neuronal responses, would be consistent with multiple functional networks underlying the respective strategies operating in parallel, guiding the behavior of the animals.

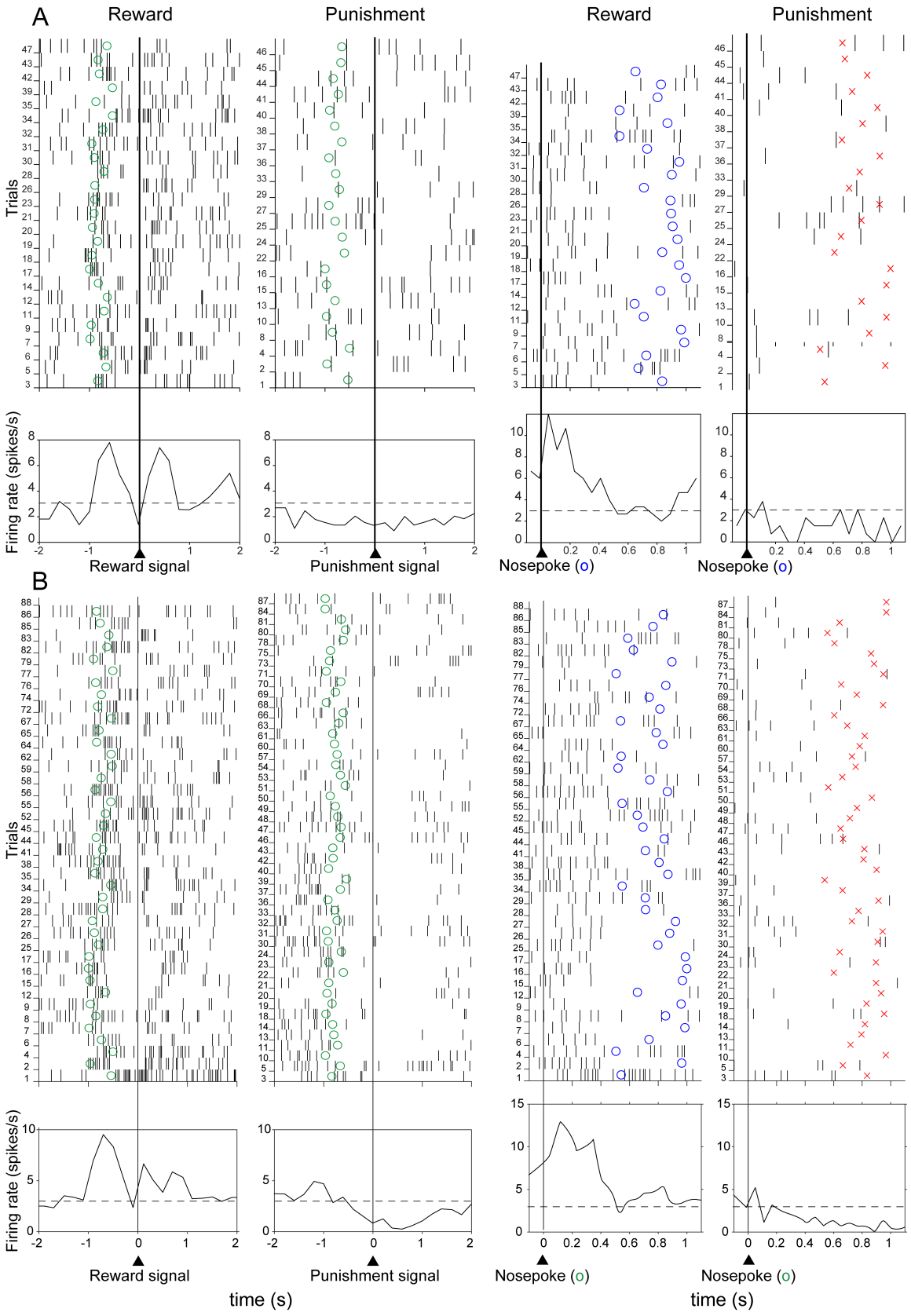


Figure 3. Examples of RP responses in sessions with performance levels of 53% (A) and 46% (B).

In both A and B, the RP response is significant in the [nose-poke, outcome signal] period, but the [cue onset, nose-poke] RP activity is only significant in A ($p<0.05$, unpaired t-test). Also,

the canonical activity increase after post-reward signals is significant in both, but activity inhibition after the post-punishment signal is significant only in B ($p<0.05$, unpaired t-test). Green circles indicate the times of nose-poke choices in left columns while at the right blue circles indicate timing of reward signals and red x's indicate punishment signals. Horizontal dashed lines in histograms indicate background firing rate prior to trial onset. These mice did not reach criterion performance in these sessions, and thus there is only one task epoch. For A, the target cue was an odor, but the mouse chose the lit port with >87% adherence on trials 3-11, and 23-45. For B, the target cue was an odor, and the mouse chose the unlit port on trials 3-12, the port emitting the non-target odor 4 on trials 18-24, alternated left and right ports on trials 42-50 (Alternation), then went to the lit port on trials 53-63 and 67-75, and to the left port on trials 78-89, again all with >87% adherence.

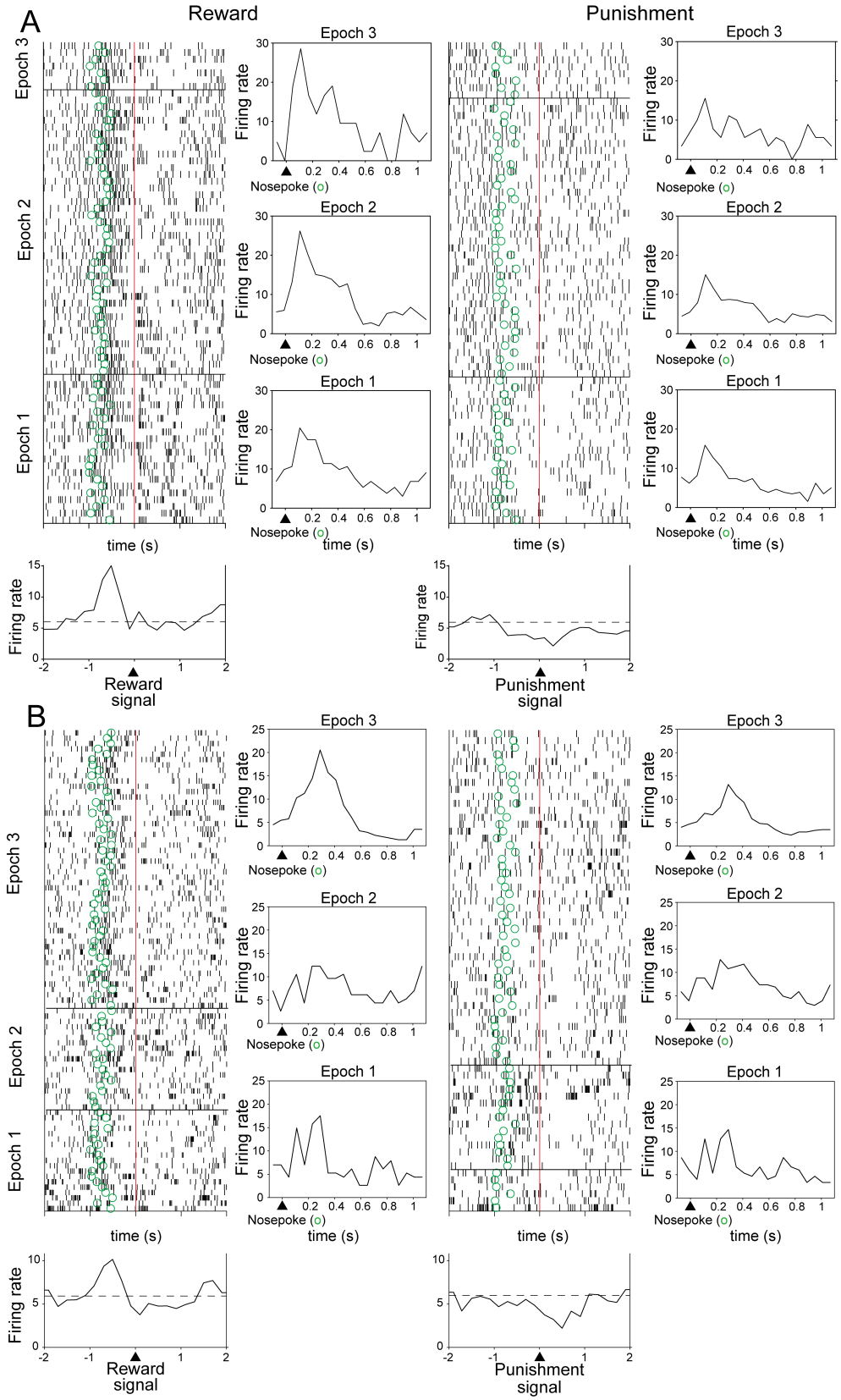


Figure 4. Examples of RP activity during the [nose-poke, outcome signal] period in two cells in different sessions.

(Same format as Figure 3.) Panels A and B show the response of two different cells. Rasters are synchronized with outcome signals while histograms in columns 2 and 4 are synchronized

with nose-poke responses (green circles in rasters). Bottom) Mean histograms from the entire session. Dashed lines indicate background firing rates. **A)** In epochs 1 and 3 the target cue was an odor, and no visual cues were presented. In epoch 2, the odor was also the target cue, but with visual cues present. RP activity was significant in all three epochs ($p<0.05$, unpaired t-test). Inhibition was significant after punishment signals in epochs 1 and 2 only ($p<0.05$, unpaired t-test), but none had significant post-reward signal excitation ($p>0.05$, unpaired t-test). Performance levels were 46%, 51% and 47% in the respective epochs. Unrewarded strategies were: trials 20-29, 93-98 and 114-131 Left port; trials 29-35, Odor 4; trials 47-56, Right port; trials 57-67 and 83-92, Unlit port; trials 72-77, Alternation. **B)** Reward predictive responses significant only in epoch 3. ($p<0.05$, unpaired t-test. Same format as A.) The inhibition after the punishment signal is significant in epoch 3 ($p<0.05$, unpaired t-test), but the firing rate was not significant higher after the reward signal ($p>0.05$, unpaired t-test). In epoch 3, unrewarded strategies were: trials 64-72 and 87-95, Unlit port; trials 75-85, 105-115 125-130, 133-150, and 153-159, Left port; Epoch 1- Visual target cue with no olfactory cues present. Epoch 2-Visual target cue with olfactory cues present. Epoch 3- Olfactory target cue with visual cues present.

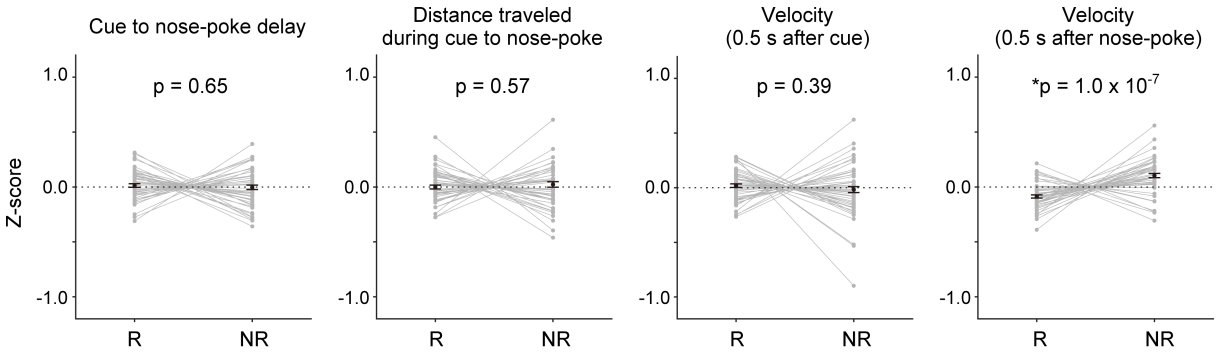
We distinguished between two levels of difficulty in the sensory discrimination tasks. When only the target cue modality was presented, the epoch was considered "simple". When both cue modalities were presented, and thus only one had to be attended to for the discrimination task, while the other had to be ignored, the epoch was considered "complex". Overall, proportions of RP responses occurred in epochs with simple and complex discriminations were not significantly different (34% vs 28%, $\chi^2(1) = 0.83$, $p = 0.36$; Supplementary Table 1). There was a greater incidence of RP responses in epochs with olfactory target cues than visual target cues (40% vs 22%; $\chi^2(1) = 10.0$, $p = 1.6 \times 10^{-3}$; see Supplementary Table 1).

Concerning anatomical localization, most recordings, and RP responses were made in the VTA (see Supplementary Figure 3D of Oberto, Matsumoto, et al., 2023) with only one epoch for a dopaminergic neuron in SNc (which was more difficult for electrode placement and the sample there is smaller) and eight epochs for OTHER SNc neurons.

Possible motor factors

We examined motor parameters in the [cue onset, nose-poke] period on rewarded and punished trials as a potential indicator of impulsivity/deliberateness of the choices. There was no significant difference for cue-to-nosepoke delay, distance traveled, or initial velocity for rewarded vs. punished trials (Supplementary Figure 4). In the period after nose-poke responses, the initial velocity was significantly greater on trials prior to punishment signals

549 was greater than prior to reward trials (Supplementary Figure 4). This raised the possibility
550 that the differential firing on rewarded vs. punished trials could be related to velocity
551 dependence of firing rate (Puryear et al, 2010). However, Pearson's correlation analyses of
552 firing rate and velocity were significant in only five of the 74 neurons ($p<0.05$, $r = -0.43$ to -
553 0.23).
554



555
556 **Supplementary figure 4. Control comparison of motor parameters in rewarded (R) and non-**
557 **rewarded (NR) trials in epochs with RP activity.**
558 Each value was z-scored based on the distribution relative to the values for the respective
559 epoch. Gray dots represent individual epoch values. Black dots and error bars indicate means
560 and SEMs. p-values are from paired t-tests. N = 74 epochs with RP responses in identified DA
561 neurons. Of 82 RP DA neurons, 8 neurons were excluded from the following analysis due to
562 technical issues in animal tracking (e.g., videos were not recorded properly).
563

Discussion

In summary, in mice training in sensory discrimination tasks in an automated double chamber, dopaminergic and unidentified (OTHER) VTA neurons fired phasically predicting reward outcome after behavioral choices, but prior to signals indicating whether the trial was to be rewarded or not. These reward-prediction responses occurred while behavioral performance was at chance and sub-criterion levels. Thus, the neurons accurately predicted the imminent reward, even though the animal's low performance level did not reflect this information. Furthermore, trial outcome signals (presented after a delay) evoked contradictory responses within the canonical "reward prediction error" (RPE) framework of interpreting DA neuronal activity during Pavlovian conditioning (Schultz, 2016). There, excitatory activity after reward signals are interpreted as the reward being *unexpected*. However, punishment signals could evoke inhibitory responses in successive trials, an indicator that an *expected* reward had been withheld. The RP responses signal reward expectation, like the post-punishment inhibition, but not the post-reward "surprise" excitatory activity. These diverse responses in DA neurons could reflect inputs from distinct brain systems that represent conflicting "belief states" (of expectation of reward; see Gershman & Uchida, 2019; Rao, 2010), as expressed in a Bayesian theoretical framework. Thus, these DA neurons signaled whether the choice was correct or not, even though this information was not exploited in the animals' performance levels. But these signals could participate in mechanisms underlying the learning process. The presence (after correct choices) and absence (on incorrect choices) of RP activity could be considered as a critic signal as postulated in reinforcement learning algorithms. The earliest appearance of RP responses occurred primarily after cue presentation when the target was visual, but after nose-pokes for odor targets. The early responses for visual targets demonstrate that the RP activity could precede the choice response, and potentially serve as an instruction signal.

While previous studies have involved learning new discriminative cues or reversal learning, the present task required shifting attention between cue modalities. Thus, the mice needed to make appropriate choices based upon the previous history of rewarded task contingencies, instead of innate unrewarded behavioral strategies or (seemingly) random choices. Because of the complex nature of these decisions, we primarily studied dopaminergic activity in VTA which receives input from the hippocampal-accumbens (limbic) pathway,

595 rather than the SNc regions whose input is from cortico-striatal pathways more associated
596 with sensorimotor functions (Haber, 2014).
597
598 **Reward prediction activity prior to criterion performance in other brain areas**
599 Schoenbaum et al (1998, 1999) found outcome predictive activity in basolateral
600 amygdala (BLA) and orbitofrontal cortex neurons of rats during pre-criterion performance in
601 a go/no-go olfactory discrimination task with a delay between onset of the behavioral choice
602 and trial outcome signals. In most BLA neurons, firing rates increased for erroneous choices
603 leading to punishment, opposite the polarity of the DA neurons here. After reversal of task
604 contingencies, but prior to criterion performance, about half of the neurons no longer
605 predicted trial outcome (Schoenbaum, et al., 1999), similar to our observations of RP activity
606 not being maintained throughout all epochs of some recording sessions. The amygdalar
607 activity also appeared at short latencies after choice onset. VTA dopaminergic neurons project
608 to interneurons in basal amygdala (Brinley-Reed & McDonald, 1999; Pinard, et al., 2008; Lutas,
609 et al., 2019; Tang et al., 2020). Furthermore, the VTA-amygdala-accumbens circuit acts as a
610 positive feedback loop to positive experiences (Sun, et al., 2021). Future work could
611 investigate potential interactions between basolateral amygdala and VTA in elaborating
612 reward predictive activity. In mouse auditory cortex neurons, Drieu, et al (2025) found a late
613 component response in neuronal activity that predicted rewards in an auditory go/no-go
614 discrimination task prior to when the animals achieved criterion performance, as found here.
615
616 **Dopaminergic activity in operant tasks**
617 While there are several studies of dopaminergic activity during instrumental learning
618 and decision-making (e.g., Bayer & Glimcher, 2005; Morris, et al., 2006; Nishino, et al., 1987;
619 Phillips, et al., 2003; Roesch, et al., 2007; Roitman, et al., 2004; Stuber, et al., 2005; Jones et
620 al., 2010), in many of these experiments, conditioned responses (CRs) and trial outcome
621 signals were concurrent, and phasic responses to the CR could not be disambiguated from
622 those to the reward signal. Studies of dopamine release in the striatum and in recordings of
623 midbrain dopaminergic neurons have typically focused on animals undergoing classical
624 conditioning or instrumental conditioning with instructed choices. Studies of instrumental
625 conditioning with free choices have provided intriguing results pointing towards models other
626 than actor/critic TD learning. Similar to the present results, Syed et al (2016) found that in an

instrumental learning task, dopamine release in striatum was greater after initiation of correct than incorrect choice responses, but prior to the reward outcome signal. When the animals made the incorrect choice, the movement dynamics were slower than on rewarded trials. However, after correct choices, the DA release dynamics were fundamentally different from the present phasic responses: the dopamine levels gradually built up over the course of several seconds to peak after the reward outcome signal. These dynamics were also observed in recordings of VTA neurons under conditions of high uncertainty (Fiorillo, et al., 2003). This type of response is generally assimilated with motivation incentive or drive for reward acquisition (Nishino, et al., 1987; Howe, et al., 2013), rather than with short-latency phasic responses related to motivation value for learning associations (Schultz et al., 1997; Wise, 2005). For example, in a task requiring self-initiated sequences of bar presses, dopamine release preceded and continued during these movements (Wassum et al., 2012). Again, the time course was over several seconds and the responses were interpreted as representing incentive motivation.

The phasic reward prediction responses here are on the same time scale as found in Pavlovian conditioning experiments, where responses to rewards or to reward-associated cues last for up to 200 ms (Schultz, 1997; Hollerman and Schultz, 1998). Lak et al (2016) distinguished a rapid time scale, at 0.1-0.2 s after cue presentation, but these corresponded to novelty responses which disappeared with familiarity. Later activity, 0.4-0.6 s after cue presentation reflected learned reward value, but appeared only after performance levels had improved. Thus, this does not correspond to the activity found here. Another response type at longer time scales is characterized by sustained gradually ramping from cue presentation until reward delivery. This is in contrast with the phasic RP activity here that immediately followed the behavioral response. For example, the RP responses here bear only superficial resemblance Goedhoop et al's (2023) recordings of DA release in the nucleus accumbens (which receives VTA inputs) in animals performing similar Pavlovian and operant conditioning tasks. Ramping of DA release occurred only in their operant task, starting with the onset of the reward-predictive cue, and leading up to the operant response. This was interpreted as anticipation of a rewarded action.

Engelhard, et al (2019) recorded DA neurons in mice performing a visual discrimination task on a virtual T-maze. During the cue presentation, but prior to choice, some neurons selectively increased their discharge rate for trials where the mice made incorrect choices.

Roesch et al (2007) found cue-triggered DA unit responses indicated the most valuable reward choice available, even on trials when the animal subsequently did not select this. After cues were turned off, but prior to movement for reward, the activity shifted to represent the actual choice. Thus, this is not similar to the observations here. Furthermore, cue selectivity developed with learning, while the responses here persisted over many trials, sometimes in sessions where learning did not occur.

Our results pose a conundrum: the neurons "know" the correct response, but this is not reflected in the animal's behavior. One possible explanation would be that the DA signal acts gradually upon the network until a critical amount of circuit changes are implemented to finally modify behavior. Here only a fraction of recorded neurons predicted reward in these sessions when the animals had not yet reached performance criterion. Perhaps further work will reveal mechanisms by which RP activity could lead to rule acquisition.

During training, behavioral choices were not random. Rather, animals performed other strategies while DA neurons predicted rewards in the visual and olfactory discrimination tasks. This could reflect the presence of multiple Bayesian belief representations which must be reconciled prior to reaching criterion performance levels. Different subsets of neurons or neuronal circuit dynamics could underlie performance of the respective strategies. Such representations were reported in mouse secondary motor (M2) cortex by Cazettes, et al. (2023) where neural ensembles simultaneously encoded multiple strategies in foraging tasks. Note that M2 projects to VTA in mice (Watabe-Uchida, et al., 2012). Functionally, dopamine release during reward prediction could prime the network for reinforcement by the subsequent reward-related dopamine release, or punishment-related dopamine absence.

References

1. Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47(1):129-41. doi: 10.1016/j.neuron.2005.05.020.
2. Berridge KC, Robinson TE. (1998) What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? Brain Res Brain Res Rev. 28(3):309-69. doi: 10.1016/s0165-0173(98)00019-8.

688 3. Birrell JM, Brown VJ (2000) Medial frontal cortex mediates perceptual attentional set
689 shifting in the rat. J Neurosci 20:4320–4324
690 4. Bissonette GB, Martins GJ, Franz TM, Harper ES, Schoenbaum G, Powell EM (2008) Doub
691 le dissociation of the effects of medial and orbital prefrontal cortical lesions on attentional
692 and affective shifts in mice. J Neurosci 28:11124–11130.
693 5. Brinley-Reed M, McDonald AJ. (1999) Evidence that dopaminergic axons provide a dense
694 innervation of specific neuronal subpopulations in the rat basolateral amygdala. Brain Res.
695 850(1-2):127-35. doi: 10.1016/s0006-8993(99)02112-5.
696 6. Cazettes F, Mazzucato L, Murakami M, Morais JP, Augusto E, Renart A, Mainen ZF (2023)
697 A reservoir of foraging decision variables in the mouse brain. Nat Neurosci 26, 840–849.
698 doi:10.1038/s41593-023-01305-8
699 7. Coddington LT, Dudman JT (2018) The timing of action determines reward prediction
700 signals in identified midbrain dopamine neurons. Nat Neurosci. 21(11):1563-1573. doi:
701 10.1038/s41593-018-0245-7.
702 8. Di Domenico D, Mapelli L. (2023) Dopaminergic modulation of prefrontal cortex inhibition.
703 Biomedicines. 11(5):1276. doi: 10.3390/biomedicines11051276.
704 9. Dias R, Robbins TW, Roberts AC (1996) Dissociation in prefrontal cortex of affective and
705 attentional shifts. Nature 380:69–72
706 10. Drieu C, Zhu Z, Wang Z, Fuller K, Wang A, Elnozahy S, Kuchibhotla K (2025) Rapid
707 emergence of latent knowledge in the sensory cortex drives learning. Nature. Mar 19.
708 doi: 10.1038/s41586-025-08730-8. Epub ahead of print.
709 11. Engelhard B, Finkelstein J, Cox J, Fleming W, Jang HJ, Ornelas S, Koay SA, Thiberge SY, Daw
710 ND, Tank DW, Witten IB (2019) Specialized coding of sensory, motor and cognitive
711 variables in VTA dopamine neurons. Nature 570:509–513.

12. Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. (2015) Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 525(7568):243-6. doi: 10.1038/nature14855.

13. Farrell K, Lak A, Saleem AB (2022) Midbrain dopamine neurons signal phasic and ramping reward prediction error during goal-directed navigation. Cell Rep. 41(2):111470. doi: 10.1016/j.celrep.2022.111470.

14. Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science. 299(5614):1898-902. doi: 10.1126/science.1077349

15. Gershman SJ, Uchida N. (2019). Believing in dopamine. Nat. Rev. Neurosci. 20, 703–714. doi.org/10.1038/s41583-019-0220-7

16. Goedhoop J, Arbab T, Willuhn I. (2023) Anticipation of appetitive operant action induces sustained dopamine release in the nucleus accumbens. J Neurosci. 43(21):3922-3932. doi: 10.1523/JNEUROSCI.1527-22.2023.

17. Haber SN. (2014) The place of dopamine in the cortico-basal ganglia circuit. Neuroscience. 282:248-57. doi: 10.1016/j.neuroscience.2014.10.008.

18. Hollerman JR, Schultz W. (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci. 1(4):304-9. doi: 10.1038/1124.

19. Houk JC, Adams JL, Barto AG. (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of Information Processing in the Basal Ganglia, eds. J. C. Houk, J. L. Davis, and D. G. Beiser. Cambridge, MA: MIT Press, p. 249–270.

734 20. Howe MW, Tierney PL, Sandberg SG, Phillips PE, Graybiel AM. (2013) Prolonged dopamine
735 signalling in striatum signals proximity and value of distant rewards. Nature.
736 29;500(7464):575-9. doi: 10.1038/nature12475.
737 21. Jones JL, Day JJ, Aragona BJ, Wheeler RA, Wightman RM, Carelli RM (2010) Basolateral
738 amygdala modulates terminal dopamine release in the nucleus accumbens and
739 conditioned responding. Biol. Psychiat. 67(8):737-44. doi:
740 10.1016/j.biopsych.2009.11.006.
741 22. Kaefer K, Nardin M, Blahna K, Csicsvari J. (2020) Replay of behavioral sequences in the
742 medial prefrontal cortex during rule switching. Neuron 106(1):154-165.e6. doi:
743 10.1016/j.neuron.2020.01.015
744 23. Khamassi, M., Lachéze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor-critic models
745 of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive*
746 *Behavior, 13*(2), 131–148. doi:10.1177/105971230501300205
747 24. Khamassi M, Peyrache A, Benchenane K, Hopkins DA, Lebas N, Douchamps V, Droulez J,
748 Battaglia FP, Wiener SI. (2024) Rat anterior cingulate neurons responsive to rule or
749 strategy changes are modulated by the hippocampal theta rhythm and sharp-wave ripples.
750 Eur J Neurosci. doi: 10.1111/ejn.16496
751 25. Kvitsiani D, Ranade S, Hangya B, Taniguchi H, Huang JZ, Kepecs A. (2013) Distinct
752 behavioural and network correlates of two interneuron types in prefrontal cortex. Nature.
753 498(7454):363-6. doi: 10.1038/nature12176.
754 26. Lak A, Stauffer WR, Schultz W. (2016) Dopamine neurons learn relative chosen value from
755 probabilistic rewards. Elife. 5:e18044. doi: 10.7554/eLife.18044.
756 27. Lapiz-Bluhm MD, Soto-Piña AE, Hensler JG, Morilak DA. (2009) Chronic intermittent cold
757 stress and serotonin depletion induce deficits of reversal learning in an attentional set-

758 shifting test in rats. Psychopharmacol (Berl). 202(1-3):329-41. doi: 10.1007/s00213-008-
759 1224-6.
760 28. Lutas A, Kucukdereli H, Alturkistani O, Carty C, Sugden AU, Fernando K, Diaz V, Flores-
761 Maldonado V, Andermann ML. (2019) State-specific gating of salient cues by midbrain
762 dopaminergic input to basal amygdala. Nat Neurosci. 22(11):1820-1833. doi:
763 10.1038/s41593-019-0506-0.
764 29. Montague PR, Dayan P, Sejnowski TJ. (1996) A framework for mesencephalic dopamine
765 systems based on predictive Hebbian learning. J Neurosci. 16(5):1936-47. doi:
766 10.1523/JNEUROSCI.16-05-01936.1996.
767 30. Nishino H, Ono T, Muramoto K, Fukuda M, Sasaki K (1987) Neuronal activity in the ventral
768 tegmental area (VTA) during motivated bar press feeding in the monkey. Brain Res.
769 413(2):302-13. doi: 10.1016/0006-8993(87)91021-3.
770 31. Oberto VJ, Matsumoto J, Pompili MN, Todorova R, Papaleo F, Nishijo H, Venance L,
771 Vandecasteele M, Wiener SI. (2023) Rhythmic oscillations in the midbrain dopaminergic
772 nuclei in mice. Front Cell Neurosci. 17:1131313. doi: 10.3389/fncel.2023.1131313.
773 32. Phillips PE, Stuber GD, Heien ML, Wightman RM, Carelli RM. (2003) Subsecond dopamine
774 release promotes cocaine seeking. Nature 422(6932):614-8. doi: 10.1038/nature01476.
775 33. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M., and Harris, K. (2016). Kilosort:
776 Realtime spike-sorting for extracellular electrophysiology with hundreds of channels.
777 bioRxiv [Preprint] doi: 10.1101/061481
778 34. Pinard CR, Muller JF, Mascagni F, McDonald AJ. (2008) Dopaminergic innervation of
779 interneurons in the rat basolateral amygdala. Neuroscience. 157(4):850-63. doi:
780 10.1016/j.neuroscience.2008.09.043.

781 35. Puryear CB, Kim MJ, Mizumori SJ. (2010) Conjunctive encoding of movement and reward
782 by ventral tegmental area neurons in the freely navigating rodent. Behav Neurosci.
783 124(2):234-47. doi: 10.1037/a0018865.
784 36. Rao RPN. (2010) Decision making under uncertainty: a neural model based on partially
785 observable Markov decision processes. Front Comput Neurosci. 4, 146.
786 37. Roesch MR, Calu DJ, Schoenbaum G. (2007) Dopamine neurons encode the better option
787 in rats deciding between differently delayed or sized rewards. Nat Neurosci. 10(12):1615-
788 24. doi: 10.1038/nn2013.
789 38. Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM (2004) Dopamine operates
790 as a subsecond modulator of food seeking. J Neurosci. 24(6):1265-71. doi:
791 10.1523/JNEUROSCI.3823-03.2004.
792 39. Scheggia D, Bebensee A, Weinberger DR, Papaleo F. (2014) The ultimate intra-/extra-
793 dimensional attentional set-shifting task for mice. Biol Psychiat. 75:660–670. doi:
794 10.1016/j.biopsych.2013.05.021
795 40. Scheggia D, Papaleo F. (2016) An operant intra-/extra-dimensional set-shift task for mice.
796 J Vis Exp. 107:e53503. doi: 10.3791/53503.
797 41. Schoenbaum G, Chiba AA, Gallagher M. (1998) Orbitofrontal cortex and basolateral
798 amygdala encode expected outcomes during learning. Nat Neurosci. 1(2):155-9. doi:
799 10.1038/407.
800 42. Schoenbaum G, Chiba AA, Gallagher M. (1999) Neural encoding in orbitofrontal cortex
801 and basolateral amygdala during olfactory discrimination learning. J Neurosci. 19(5):1876-
802 84. doi: 10.1523/JNEUROSCI.19-05-01876.1999.
803 43. Schultz W. (2016) Dopamine reward prediction-error signalling: a two-component
804 response. Nat Rev Neurosci. 2016 17(3):183-95. doi: 10.1038/nrn.2015.26.

805 44. Schultz W, Dayan P, Montague PR. (1997) A neural substrate of prediction and reward.
806 Science. 275(5306):1593-9. doi: 10.1126/science.275.5306.1593.
807 45. Stuber GD, Roitman MF, Phillips PE, Carelli RM, Wightman RM. (2005) Rapid dopamine
808 signaling in the nucleus accumbens during contingent and noncontingent cocaine
809 administration. Neuropsychopharmacology. 30(5):853-63. doi: 10.1038/sj.npp.1300619.
810 46. Syed EC, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME. (2016) Action initiation
811 shapes mesolimbic dopamine encoding of future rewards. Nat Neurosci. 19(1):34-6. doi:
812 10.1038/nn.4187.
813 47. Tait DS, Phillips JM, Blackwell AD, Brown VJ. (2017) Effects of lesions of the subthalamic
814 nucleus/zona incerta area and dorsomedial striatum on attentional set-shifting in the rat.
815 Neurosci. 345:287-296. doi: 10.1016/j.neuroscience.2016.08.008.
816 48. Tang W, Kochubey O, Kintscher M, Schneggenburger R. (2020) A VTA to basal amygdala
817 dopamine projection contributes to signal salient somatosensory events during fear
818 learning. J Neurosci. 40(20):3969-3980. doi: 10.1523/JNEUROSCI.1796-19.2020.
819 49. Ungless MA, Grace AA. (2012) Are you or aren't you? Challenges associated with
820 physiologically identifying dopamine neurons. Trends Neurosci. 35(7):422-30. doi:
821 10.1016/j.tins.2012.02.003.
822 50. Vander Weele CM, Siciliano CA, Tye KM. (2019) Dopamine tunes prefrontal outputs to
823 orchestrate aversive processing. Brain Res. 1713:16-31. doi:
824 10.1016/j.brainres.2018.11.044.
825 51. Wassum KM, Ostlund SB, Maidment NT. (2012) Phasic mesolimbic dopamine signaling
826 precedes and predicts performance of a self-initiated action sequence task. Biol
827 Psychiatry. 71(10):846-54. doi: 10.1016/j.biopsych.2011.12.019.

828 52. Watabe-Uchida M, Zhu L, Ogawa SK, Vamanrao A, Uchida N. (2012) Whole-brain mapping
829 of direct inputs to midbrain dopamine neurons. Neuron. 74(5):858-73. doi:
830 10.1016/j.neuron.2012.03.017.
831 53. Wise RA. (2005) Forebrain substrates of reward and motivation. J Comp Neurol.
832 493(1):115-21. doi: 10.1002/cne.20689.

833

Author contributions

JM and SIW designed and performed the experiments and obtained funding support. MNP performed experiments. JM designed the carrier for optic fibers and octrode driver assemblies, constructed the experimental apparatus, the optrodes and optogenetic stimulation apparatus, informatics control, and data acquisition systems with support from HN. MV and LV maintained and genotyped the mouse line and guided immunohistochemical processing. FP guided adaptation of the maze and behavioral protocols for training and recording. JM, VJO, and SIW analyzed the data with support from RT and HN, and wrote the manuscript. All authors approved of the manuscript.

Funding

Grants from Uehara Memorial Foundation (to JM), the Takeda Science Foundation (to JM and HN), the Labex Memolife, Fondation pour la Recherche Médical, Fondation Bettencourt Schueller, and International Research Laboratory DALoops (to SW) provided support.

Ethics

All procedures were in accord with local (autorisation d'experimenter no. 75-1328-R; Comite d'Ethique pour l'Experimentation Animale no. 59, dossier 2012-0007) and international (European Directive 2010/63/EU; US National Institutes of Health guidelines) standards and legal regulations regarding the use and care of animals.