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

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, Toyama, Japan

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

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

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

* Corresponding author

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

Abstract

In the canonical interpretation of phasic activation of dopaminergic neurons during Pavlovian conditioning, initially cell firing is triggered by unexpected rewards. Upon learning, activation instead follows the reward-predictive conditioned stimulus. When expected rewards are withheld, firing is inhibited. Here, we recorded optogenetically identified dopaminergic neurons of ventral tegmental area (VTA) in mice training in successive operant sensory discrimination tasks. A delay was imposed between nose-poke choices and trial outcome signals (reward or punishment). While animals were still performing at sub-criterion levels in the task, firing increased after correct choices, but prior to trial outcome signals. Thus, the neurons predicted whether choices would be rewarded, despite the animals' poor behavioral performance. Surprisingly, these neurons also fired after reward

delivery, as if the rewards had been unexpected, but the cells were inhibited after punishment signals, as if the reward had been expected after all. These inconsistencies suggest extension of theoretical formulations of dopaminergic neuronal activity: it would embody multiple roles in temporal difference learning and actor-critic models. Furthermore, during chance and sub-criterion performance levels during task training, the mice performed other task strategies (e.g., alternation and spatial persistence) which did not reliably elicit rewards, again while these neurons predicted the correct choice. The reward prediction activity of these neurons could serve as critic signal for the preceding choice. These finding are consistent with the notion that multiple Bayesian belief representations must be reconciled prior to reaching criterion performance levels.

Introduction

Adaptation for survival requires associating predictive cues with appropriate behavioral responses to favor reward and to avoid aversive outcomes. Dopaminergic activity is implicated in the modification of neural circuits making such associations. In animals undergoing Pavlovian conditioning, dopaminergic neurons fire phasically after unexpected rewards. The DA neuron responses to rewards persist on trials 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 action selection (e.g., Khamassi, et al., 2005). These formulations originated in studies of Pavlovian conditioning, and their

extension to instrumental conditioning still requires further elaboration regarding their role in developing appropriate actions to optimize rewards and minimize aversive outcomes.

For example, in instrumental tasks, if the animal has learned that performing the appropriate action, the CR, will reliably predict reward, then, this might also be expected to trigger a short latency phasic response, as observed in Pavlovian conditioning after learning when the CS is presented. Indeed, Benchenane, et al. (2010) observed neural correlates of dopaminergic activity in rats at the decision point of a Y-maze, but only once they began performing at criterion levels in spatial or visual discrimination tasks. However, to our knowledge, dopaminergic neuron activity during conditioned responses has not been observed directly.

To test for phasic activity in dopaminergic neurons following CR's, we recorded mice as they trained in odor discrimination and visual discrimination tasks. The mice were a mutant strain, permitting optogenetic identification of dopaminergic neurons with light stimulation.

As expected, we found phasic dopaminergic neuron activity in the 0.5-1.0 s interval between cue onset and trial outcome signals. Interestingly they discharged following correct, but not incorrect choices. Thus, these responses signaled "reward prediction (RP)". However, surprisingly, this activity appeared while the animals were still learning the task, and had not yet achieved criterion performance levels, and thus could serve as a critic signal. Most of these same neurons also had typical excitatory responses to rewards (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 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 CR is evaluated by the dopaminergic critic prior to trial outcome. These multiple responses could reflect multiple Bayesian belief representations which must be reconciled prior to reaching criterion performance levels. Furthermore, the mice often performed other non-optimal strategies during RP activity, suggesting multiple parallel coding modules elsewhere in the brain. 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.

Results

Mice with optogenetically responsive dopaminergic neurons were first trained in a visual discrimination task in a double chamber apparatus (Figure 1). At the beginning of each trial, a central barrier was lowered. Once the mouse crossed a photodetector to enter the other chamber, cues for the visual discrimination task were presented simultaneously at the left and right nose-poke ports. Nose-poke choices were followed by a delay at a variable latency ranging from 0.5 to 1.0 s, and then triggered either the audible food pellet dispenser (for correct choices), or, alternatively, a punishment (5 or 7 seconds with room lights on coupled with a delay for barrier removal and the next trial).

Mice were then implanted with multiple electrodes and equipped for optical stimulation. An olfactory discrimination task was then introduced in the same maze, with the odors emanating from the nose-poke ports. Once the mouse achieved criterion performance (6 successive rewarded trials, or 8 out of 10), the task was changed: if only one cue modality had been presented, the other was added without changing the reward contingency. If the two modalities had been presented, the reward contingency was shifted to the other one. Thus, the animals were continuously challenged to adhere to the currently rewarded rule in the presence or absence of other modality cues, or to shift to the other sensory discrimination task.

In 159 recording sessions in four mice, of 659 recorded neurons, 171 were optogenetically identified as dopaminergic (DA) neurons, and 113 were fast-spiking interneurons (FSI), while 375 others could not be identified. The latter are referred to as “OTHER” neurons (cf., Oberto, Matsumoto, et al., 2023 for criteria). While mice were training in the sensory discrimination tasks, but performing at sub-criterion levels (56±1%, mean ± SEM; Supplementary Figure 1), optogenetically identified DA neurons fired after the mice made nosepoke choices, but, importantly, prior to the signals indicating whether the choice would be rewarded or punished (Figure 2, Supplementary Figure 2 for OTHER neurons). A greater number of neurons significantly increased firing when predicting rewards than punishments (Supplementary Figure 3). In the RP cells, when the target (currently rewarded) cue was

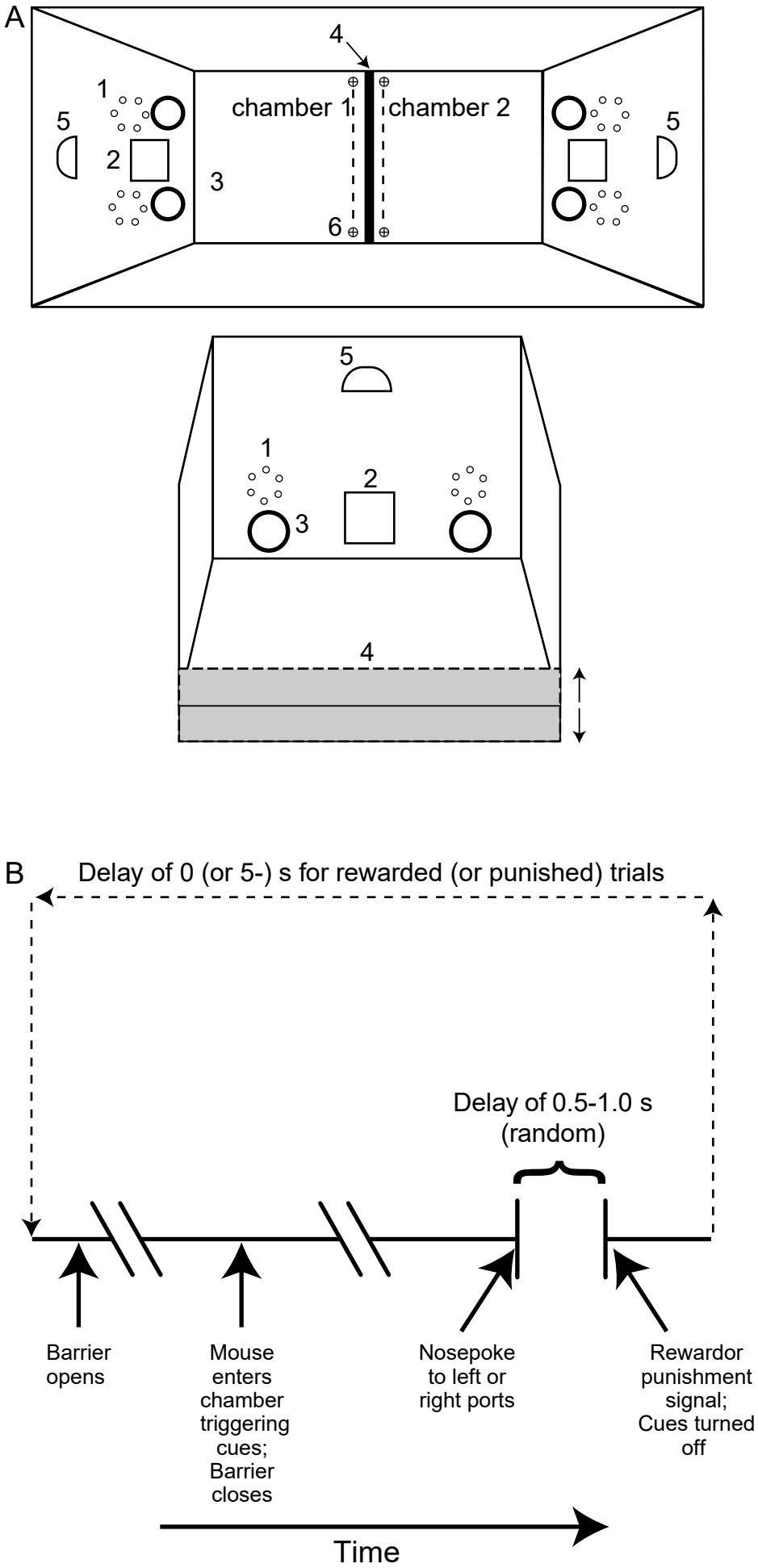
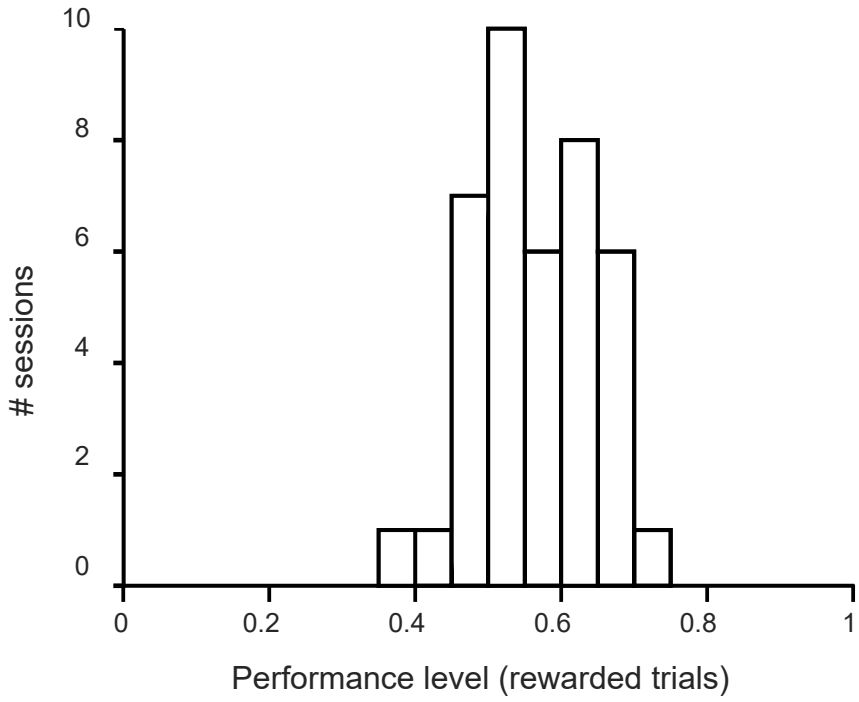


Figure 1. The behavioral apparatus and task (after Scheggia, et al., 2014). A) Top) Overhead view. 1- Hexagonal LED array; 2- Reward dispenser; 3- Nose-poke port; 4- Moveable barrier; 5-House light; 6- Photobeam. Bottom) Frontal view of one chamber (photobeam not shown here). B) The automated task sequence.

visual, the earliest responses began shortly after the discriminative cue was presented (Figure 2D left; significant divergence in cumulative sum traces is indicated by the horizontal line). When target cues were odors, the mean difference in firing rate between reward and punishment signals became significant only after the nose-poke choice responses (Figure 2D right, horizontal line). In neither case, did the activity continue to ramp up until trial outcome signals were presented (Figure 2C right panel, Supplementary Figure 2, right panel). Thus, they do not correspond to ramping responses previously found in dopaminergic neurons prior to rewards or reward predicting signals (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 Figure 1. Distribution of mean performance levels (proportion of trials that were rewarded) in sessions with reward-predictive responses.

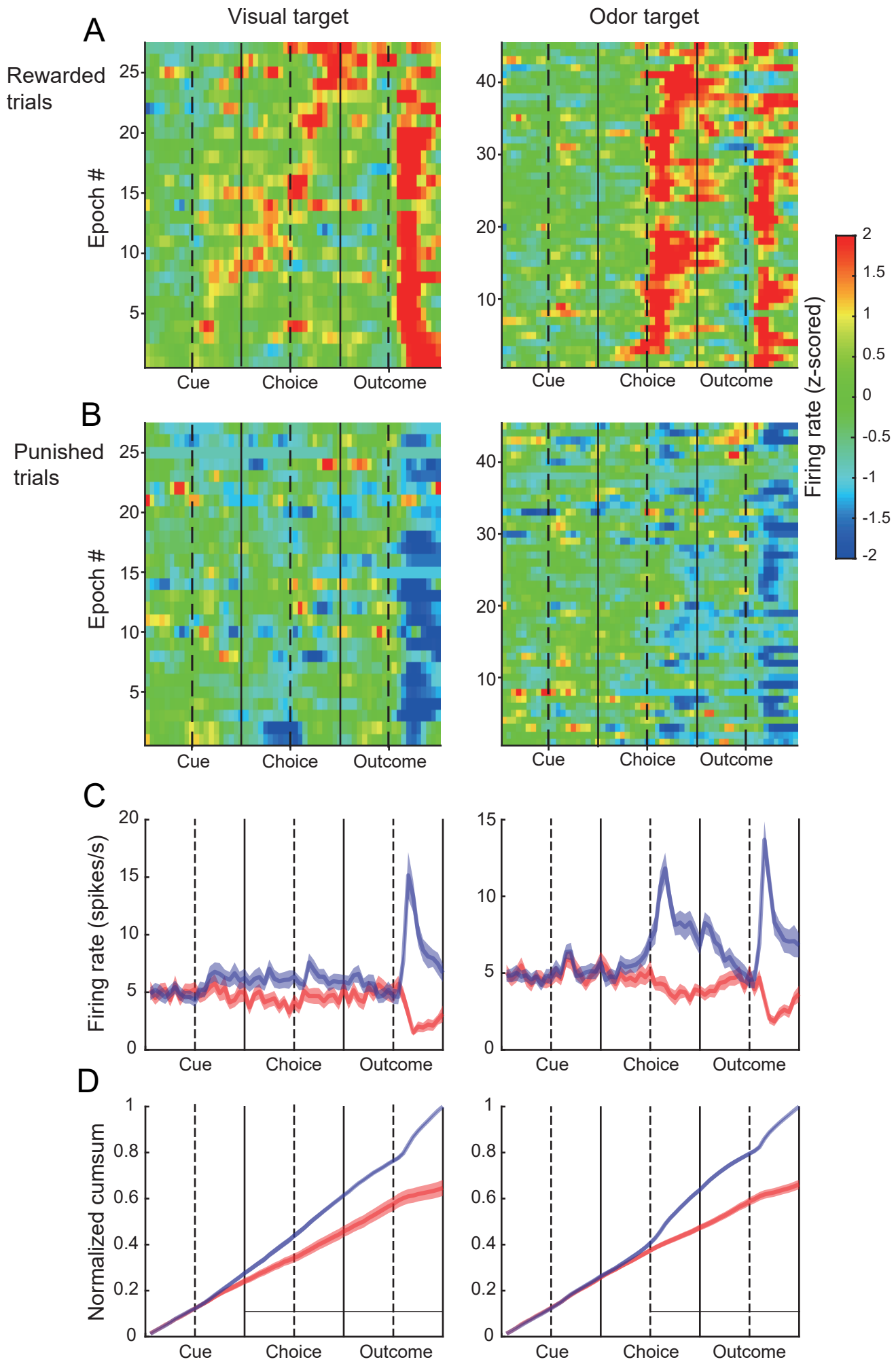
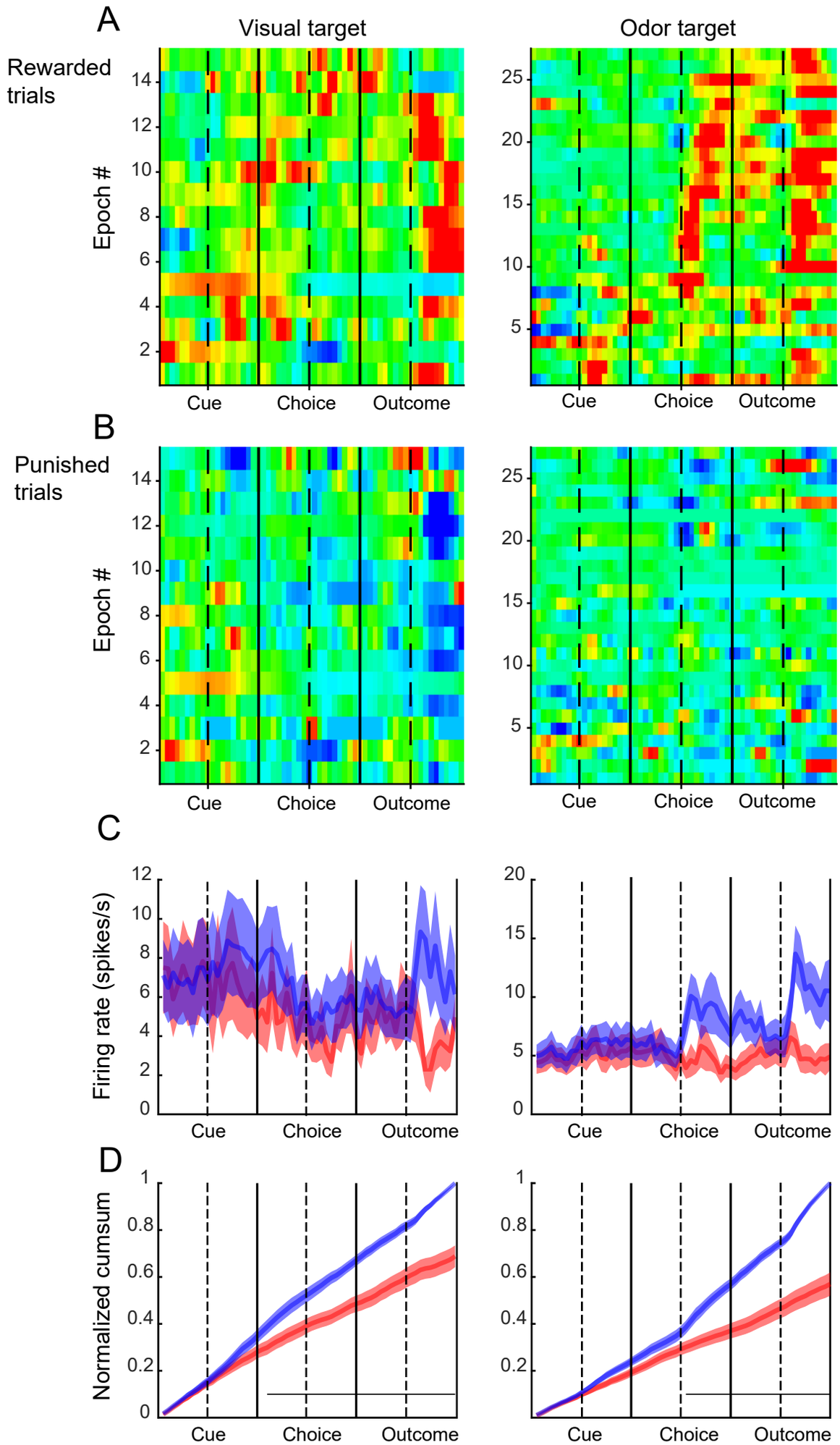
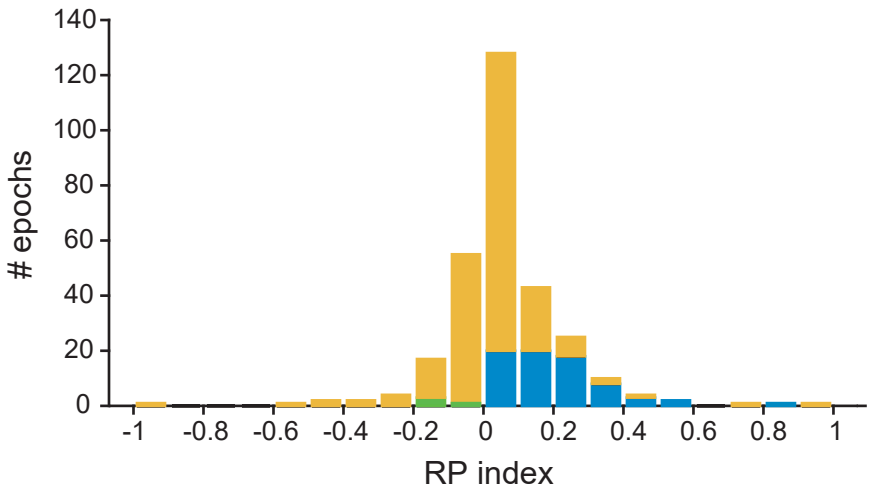

A
Visual target
Odor target
Rewarded trials
Epoch #
25
20
15
10
5
40
30
20
10
Cue
Choice
Outcome
B
Punished trials
Epoch #
Firing rate (z-scored)
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
C
Firing rate (spikes/s)
20
15
10
5
0
D
Normalized cumsum
1
0.8
0.6
0.4
0.2
0

Figure 2. Reward predictive responses in representative optogenetically identified midbrain dopaminergic neurons. A, B) 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 within the figures represents firing during a single epoch of a neuron with a significant RP response. (Data were analyzed separately for trials for each different task condition, called an epoch here, within the sessions.) C) Histograms summing raster data of A and B. Blue traces are rewarded trials and red traces are punished trials. D) Cumulative sums of histograms in C show the timing of appearance of firing rate differences between rewarded and punished trials. Horizontal bars (lower right) show where the curves were significantly different ($p<.05$, resampling method of Fujisawa et al, 2008).



Supplementary Figure 2. OTHER neurons with RP responses. Same format as Figure 2.



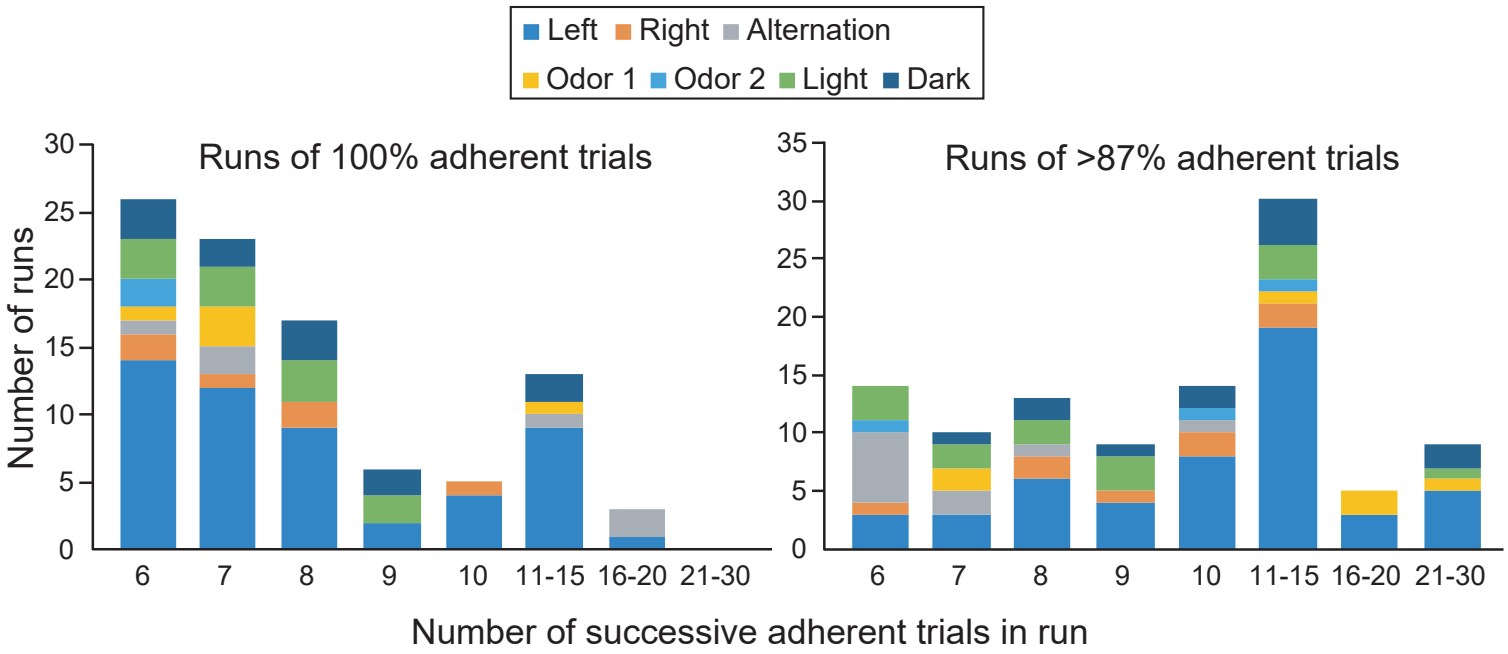
Supplementary Figure 3. The distribution of reward prediction indices of the dopaminergic neurons over all epochs. This index is calculated from the normalized firing rates in the period between nose-poke choice and reward as (rewarded/punished)/(rewarded+punished). Blue: R>NR; Red: NR>R; Yellow: n.s.

Figures 3 and 4 show examples of these RP responses in optogenetically identified dopaminergic neurons. The neurons of Figures 3A and B were recorded in sessions where the animals did not reach criterion performance (overall 53% and 47% correct, respectively). In addition to their prominent RP responses, they also increased firing rate by at least 50% (relative to baseline) after reward onset, as previously reported in Pavlovian conditioning prior to acquisition (Schultz, et al., 1997). This reward response occurred in the vast majority of the RP neurons (Figure 2 and Supplementary Figure 2). But, another RPE response, inhibition after the punishment onset (defined here as a mean firing rate reduction by at least 33% relative to baseline in at least one histogram bin), is evident in Figures 2 and 3, and Supplementary Figure 2. In the literature, this response 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 absence of RP activity on these trials would indicate that no reward was expected. Note that when this inhibition occurred (in 90% of the neurons), it only lasted at most for 1 s after punishment onset, even though the punishment period lasted 5 to 7 s.

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%). (The reason for RP responses in only a subset of epochs may be related to the identity of the target cue and complexity of the task, discussed below.) Thus, below, the incidence of RP responses is tallied by task epochs rather than over whole sessions (which could dilute significant effects).

Interestingly, 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, currently unrewarded odor or lit/unlit arm). In epochs with RP responses, 37 ± 3% (mean ± SEM; range 0-72%) of trials were performed according to strategies other than the current rule (n=33 epochs; see Methods). The numbers of successive trials in runs with adherence to non-rewarded strategies is shown in Supplementary Figure 4. In Figures 3A and B and 4 A and B, the incidences of unrewarded strategies in RP epochs were 63%, 37%, 33% and 48% respectively. 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.

The incidence of RP responses was greater in DA neurons (25%) than FSI (8%; chi-squared=25.3, $p<1e-5$) or OTHER neurons (9%; chi-squared=47.4, $p<1e-5$). The results reported below focus on the DA neurons, because of their theoretical interest.



Supplementary Figure 4. Incidences of the lengths of the 301 runs of successive trials adherent to strategies other than the current rewarded rule for 33 epochs when 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 2243 trials. (See Methods for further details.) In Figure 3A, the target cue was odor 1, and the mouse (#2) chose the lit port ("light strategy") with >87% adherence on trials 3-11, and 23-45. In Figure 3B, the target cue was odor 1, and the mouse (#1) chose the unlit port on trials 3-12 ("dark strategy"), the port emitting odor 2 on trials 18-24, alternated left and right ports on trials 42-50 ("alternation strategy"), went to the lit port on trials 53-63 and 67-75, and to the left port on trials 78-89 ("left strategy"), again all with >87% adherence.

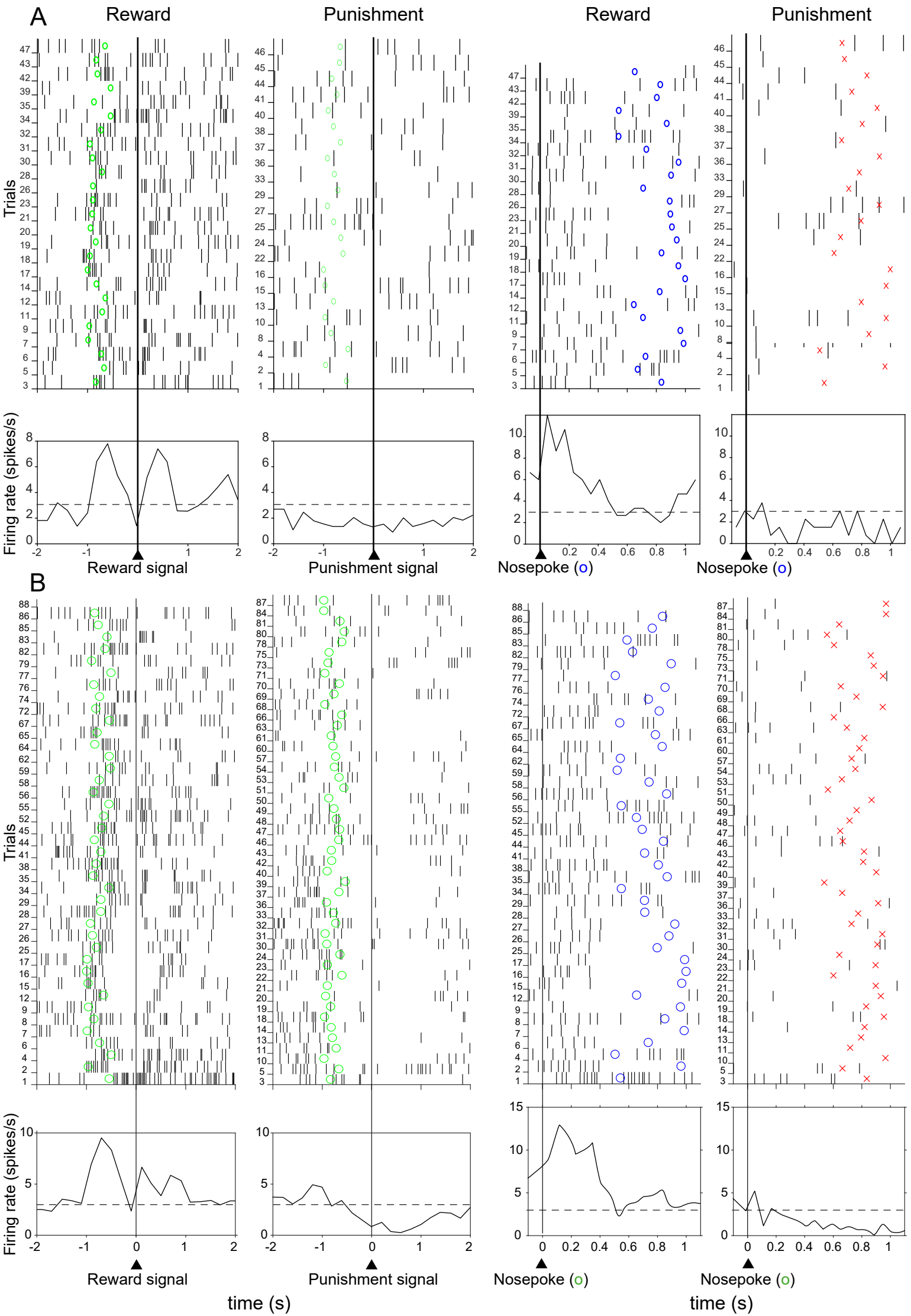

A
Reward
Punishment
Reward
Punishment
Trials
Firing rate (spikes/s)
Reward signal
Punishment signal
Nosepoke (o)
Nosepoke (o)
B
Trials
Firing rate (spikes/s)
Reward signal
Punishment signal
Nosepoke (o)
Nosepoke (o)
time (s)
time (s)

Figure 3. Examples of reward predictive responses in two different sessions. Reward prediction and post-reward signal activities occur on the same trial (e.g., trials 5 and 6 in A). Also, post-reward signal activity and post-punishment signal inhibition occur in successive trials (e.g., trials 4 and 5 in A). The first two columns are synchronized with the reward and punishment signals. Green circles in the raster displays indicate the timing of nose-poke choices. The last two columns show the same data synchronized with the nose-poke responses. 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. Mice did not reach criterion performance in these sessions, and thus only one task was presented.

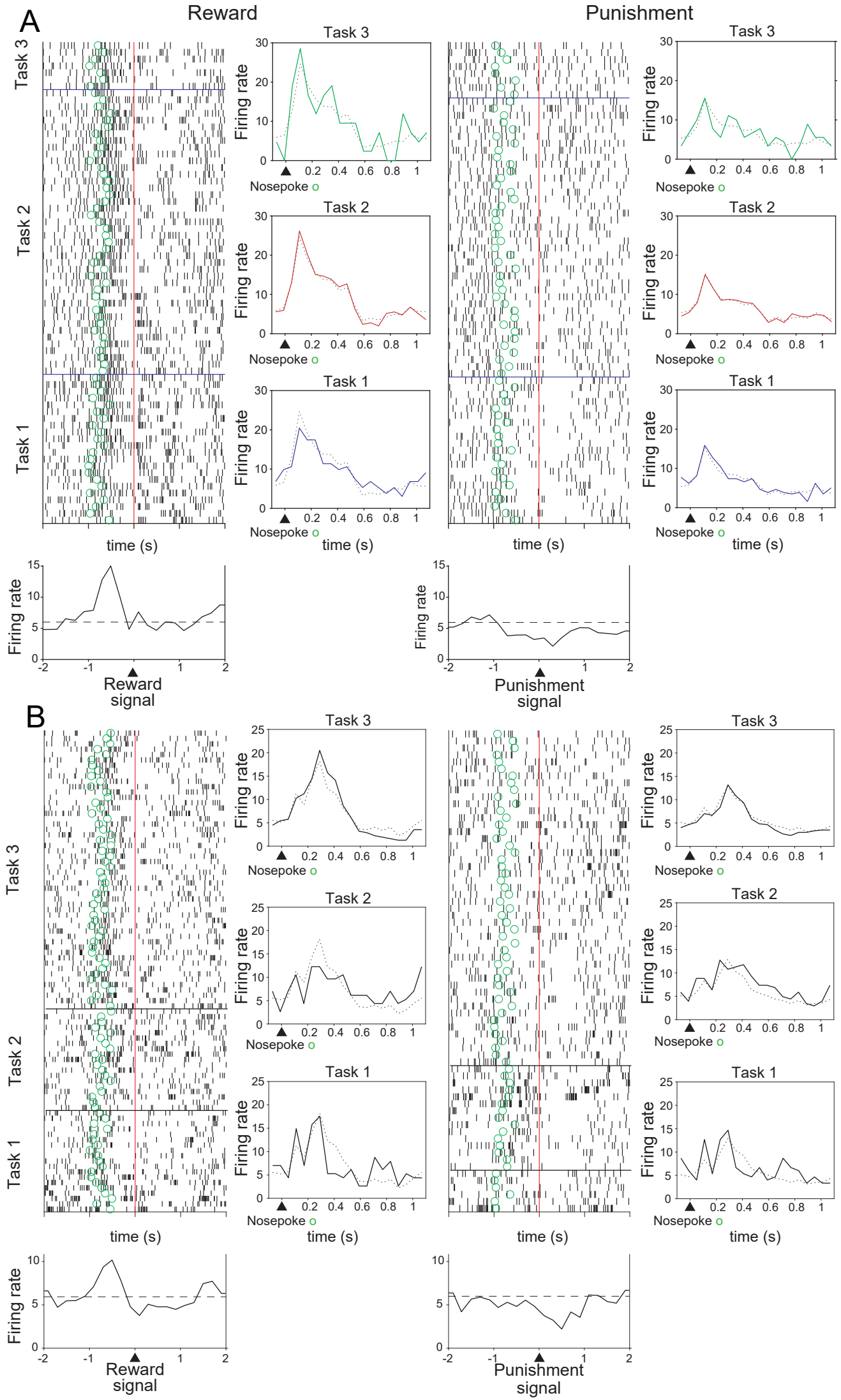

A
Reward
Punishment
Task 3
Task 2
Task 1
Firing rate
Nosepoke o
time (s)
Reward signal
Punishment signal
B

Figure 4. Significant reward prediction responses under different task conditions (1-3). Histograms in columns 2 and 4 are synchronized with nose-poke responses. Dotted traces reproduce the whole session histograms from bottom. A) Tasks 1 and 3- Olfactory target cue with no visual cues. Task 2 also presented the olfactory target cue, but with visual cues present. B) Reward predictive responses significant only in task #3. (Same format as A.) Note also the inhibition after the punishment signal. Task 1- Visual target cue with no olfactory cues present. Task 2-Visual target cue with olfactory cues present. Task 3- Olfactory target cue with visual cues present. Same format as Figure 3.

Relative incidence of different types of RP responses in optogenetically identified DA neurons

We distinguished two types of sensory discrimination trials. When only the target cue modality was presented, the trial was considered "simple". When both cue modalities were presented, one had to be attended to for the discrimination task, while the other had to be ignored, and thus these trials are considered "complex". Overall, a greater proportions of RP responses occurred in epochs with complex rather than simple discriminations (33% vs 19%, p=.00084, chi-squared= 6.9; see Table 1). Comparing between epochs when the target (rewarded) cue modality was olfactory or visual, there was a greater incidence for olfactory (34% vs 16%; see Table 1; p=.0001, chi-square=15.0). Recall that prior to surgery for implanting headstages, mice had been trained in the visual discrimination task only, with no olfactory cues. So, they had more experience with the visual discrimination, and the olfactory discrimination may have been more challenging. Combining these distinctions, the greatest incidence of RP occurred in epochs with a complex discrimination and an olfactory target cue (53%). This warrants further investigation in future studies. Very few epochs had responses that were greater for the un-rewarded cue than the rewarded cue (Table 1), and thus were not included in these tallies.

Concerning anatomical localization, most recordings, and RP responses were made in the VTA with only one epoch for a dopaminergic neuron in SNc (which was more difficult for electrode placement) and eight epochs for OTHER SNc neurons.

Possible motor factors

We examined motor parameters in the period between cue onset and the nose-poke decisions on rewarded and punished trials as a potential indicator of impulsivity of the choices. There was no significant difference for cue-to-nosepoke delay, distance travelled, or initial velocity for rewarded vs. punished trials (pair-wise t-tests by epoch; p=0.44, 0.91, 0.54). In the period when the RP activity occurred, i.e., after nose-poke responses, the initial

velocity was significantly greater on trials prior to punishment signals was greater than prior to reward trials (p=0.002). This raised the possibility that the differential firing on rewarded vs punished trials could be related to velocity dependence of firing rate (Puryear et al, 2010). However, regression analyses of firing rate and velocity were significant in only four of the 67 neurons (p<0.05, with 3 increasing and 1 decreasing).

	Light target cue			Odor target cue			All		
	simple	complex	Total	simple	complex	Total	simple	complex	Total
R>NR	11	16	27	28	17	45	39	33	72
NR<R	0	1	1	3	0	3	3	1	4
n.s.	93	51	144	68	15	83	161	66	227
Total	104	68	172	99	32	131	203	100	303
% R>NR	11%	24%	16%	28%	53%	34%	19%	33%	24%

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

Discussion

In summary, in mice training in sensory discrimination tasks in an automated double chamber, dopaminergic neurons fired phasically following correct nose-poke responses, but prior to signals independently of 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, following a delay after the nose-poke responses, trial outcome signals evoked contradictory responses within the "reward prediction error" (RPE) framework of interpreting dopaminergic neuronal activity during Pavlovian conditioning (Schultz, 2016). There, excitatory activity after reward signals are interpreted as the reward being unexpected. However, punishment signals evoked inhibitory responses, an indicator that an *expected* reward had been withheld. These diverse responses in dopaminergic 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.

Here, individual neurons displayed responses associated with rewards being both expected (reward prediction and post-punishment inhibition) and unexpected (post-reward firing rate increase) on successive trials. Crucially, the reward predictive firing occurred while the animals’ behavior was well below criterion performance, and the animals were learning the stimulus-response-reward contingencies. Thus, these neurons signaled whether the choice was correct or not, even though this information was not employed in the animal’s choices as reflected by their performance levels, and thus could have participated in the learning process. The presence (after correct choices) and absence (on incorrect choices) could be considered as a critic signal as postulated in reinforcement learning algorithms.

RP responses were more prevalent in epochs when both cue modalities were present (“complex” trials), and when the target cue was olfactory. These may have been more challenging to the mice than other trial types. The earliest appearance of reward prediction responses after cue presentation when the target was visual, but after nose-pokes for slower processing of the cues. The early responses for visual targets demonstrate that the reward prediction activity could precede the choice response, and serve as an instruction signal.

Analogous to the CS in Pavlovian conditioning, here, in instrumental conditioning, the CS (for some visual target epochs), and the CR (nose-poke choice for the other epochs) that triggered dopaminergic neuron firing, but this occurred while the animal was still learning, not after acquisition. While there are several studies of dopaminergic activity during instrumental learning and decision-making (Bayer & Glimcher, 2005; Morris, et al., 2006; Nishino, et al., 1987; Phillips, et al., 2003; Roesch, et al., 2007; Roitman, et al., 2004; Stuber, et al., 2005; Jones et al., 2010), in these experiments, CRs and trial outcome signals were concurrent, and phasic responses to the CR could not be disambiguated from those to the reward signal.

Comparison with published responses in operant tasks

Studies of dopamine release in the striatum and in recordings of midbrain dopaminergic neurons have typically focused on animals undergoing classical conditioning or instrumental conditioning with instructed choices. Studies of instrumental conditioning with free choices

have provided intriguing results pointing towards models other 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 (Syed et al, 2016). 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. In a study of non-human primates performing a two-arm bandit task, Morris et al. (2006) observed that population activity of DA neurons corresponds to choice behaviors. However, the latter held only for instructed choice trials, not for free choice decision trials, as was the case here. They concluded that dopamine neurons receive information about upcoming actions, and argue for including Q-learning into algorithms describing this activity. A challenge to a role for RPE in learning comes from Coddington and Dudman's (2018) study of trace conditioning in mice. There, the RPE correlates were found in VTA and SNc dopaminergic neurons, but these only appeared after learning. Also, Lak et al. studied monkeys learning the values of probabilistic rewards. Even though behavior indicated that acquisition occurred within a few trials, DA activity corresponding to learning (a late response appearing 0.4-0.65 s after the cue) only developed later, after the 4th trial.

Timing in relation to previously reported DA responses

The phasic CR-triggered reward prediction responses here are on the same time scale as found in Pavlovian conditioning, 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 responses here that immediately followed the CR. 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.

Reward prediction, and mismatch with behavior

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 while the animals had not yet reached performance criterion.

During training, behavioral choices were not random. Rather, animals performed other strategies while DA neurons predicted rewards in the visual and olfactory discrimination task. 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.

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.

Author contributions

JM and SIW designed and performed the experiments. MNP performed experiments. JM 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) and the Takeda Science Foundation (to JM and HN). The Labex Memolife, Fondation pour la Recherche Médical, Fondation Bettencourt Schueller, and International Research Laboratory DALoops provided support.

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, Dr. Guillaume Dugué and NeuroFabLab for facilitating with 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.

References

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.

Benchenane K, Peyrache A, Khamassi M, Tierney PL, Gioanni Y, Battaglia FP, Wiener SI. (2010) Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. Neuron (2010) 66(6):921-36. doi: 10.1016/j.neuron.2010.05.013.

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.

Coddington LT, Dudman JT (2018) The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci. 21(11):1563-1573. doi: 10.1038/s41593-018-0245-7.

Engelhard B, Finkelstein J, Cox J, Fleming W, Jang HJ, Ornelas S, Koay SA, Thiberge SY, Daw ND, Tank DW, Witten IB (2019) Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature 570:509–513.

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.

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

Fujisawa S, Amarasingham A, Harrison M, Buzsáki G (2008) Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat Neurosci. 11, 823–833. doi.org/10.1038/nn.2134

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

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.

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.

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.

Howe MW, Tierney PL, Sandberg SG, Phillips PE, Graybiel AM. (2013) Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature. 29;500(7464):575-9. doi: 10.1038/nature12475.

Jones JL, Day JJ, Aragona BJ, Wheeler RA, Wightman RM, Carelli RM (2010) Basolateral amygdala modulates terminal dopamine release in the nucleus accumbens and conditioned responding. Biol. Psychiat. 67(8):737-44. doi: 10.1016/j.biopsych.2009.11.006.

Kaefer K, Nardin M, Blahna K, Csicsvari J. (2020) Replay of behavioral sequences in the medial prefrontal cortex during rule switching. Neuron 106(1):154-165.e6. doi: 10.1016/j.neuron.2020.01.015

Khamassi, M., Lachéze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior, 13*(2), 131–148. https://doi.org/10.1177/105971230501300205

Lak A, Stauffer WR, Schultz W. (2016) Dopamine neurons learn relative chosen value from probabilistic rewards. Elife. 5:e18044. doi: 10.7554/eLife.18044.

Lapiz-Bluhm MD, Soto-Piña AE, Hensler JG, Morilak DA. (2009) Chronic intermittent cold stress and serotonin depletion induce deficits of reversal learning in an attentional set-shifting test in rats. Psychopharmacol (Berl) 202(1-3):329-41. doi: 10.1007/s00213-008-1224-6.

Montague PR, Dayan P, Sejnowski TJ. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 16(5):1936-47. doi: 10.1523/JNEUROSCI.16-05-01936.1996.

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci. 9(8):1057-63. doi: 10.1038/nn1743.

Nishino H, Ono T, Muramoto K, Fukuda M, Sasaki K (1987) Neuronal activity in the ventral tegmental area (VTA) during motivated bar press feeding in the monkey. Brain Res. 413(2):302-13. doi: 10.1016/0006-8993(87)91021-3.

Oberto VJ, Matsumoto J, Pompili MN, Todorova R, Papaleo F, Nishijo H, Venance L, Vandecasteele M, Wiener SI. (2023) Rhythmic oscillations in the midbrain dopaminergic nuclei in mice. Front Cell Neurosci. 17:1131313. doi: 10.3389/fncel.2023.1131313.

Phillips PE, Stuber GD, Heien ML, Wightman RM, Carelli RM. (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422(6932):614-8. doi: 10.1038/nature01476.

Puryear CB, Kim MJ, Mizumori SJ. (2010) Conjunctive encoding of movement and reward by ventral tegmental area neurons in the freely navigating rodent. Behav Neurosci. 124(2):234-47. doi: 10.1037/a0018865.

Rao RPN. (2010) Decision making under uncertainty: a neural model based on partially observable Markov decision processes. Front Comput Neurosci. 4, 146.

Roesch MR, Calu DJ, Schoenbaum G. (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci. 10(12):1615-24. doi: 10.1038/nn2013.

Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM (2004) Dopamine operates as a subsecond modulator of food seeking. J Neurosci. 24(6):1265-71. doi: 10.1523/JNEUROSCI.3823-03.2004.

Scheggia D, Bebensee A, Weinberger DR, Papaleo F. (2014) The ultimate intra-/extra-dimensional attentional set-shifting task for mice. Biol Psychiat. 75:660–670. doi: 10.1016/j.biopsych.2013.05.021

Scheggia D, Papaleo F. (2016) An operant intra-/extra-dimensional set-shift task for mice. J Vis Exp. 107:e53503. doi: 10.3791/53503.

Schultz W. (2016) Dopamine reward prediction-error signalling: a two-component response. Nat Rev Neurosci. 2016 17(3):183-95. doi: 10.1038/nrn.2015.26.

Schultz W, Dayan P, Montague PR. (1997) A neural substrate of prediction and reward. Science. 275(5306):1593-9. doi: 10.1126/science.275.5306.1593.

Stuber GD, Roitman MF, Phillips PE, Carelli RM, Wightman RM. (2005) Rapid dopamine signaling in the nucleus accumbens during contingent and noncontingent cocaine administration. Neuropsychopharmacology. 30(5):853-63. doi: 10.1038/sj.npp.1300619.

Syed EC, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME. (2016) Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat Neurosci. 19(1):34-6. doi: 10.1038/nn.4187.

Tait DS, Phillips JM, Blackwell AD, Brown VJ. (2017) Effects of lesions of the subthalamic nucleus/zona incerta area and dorsomedial striatum on attentional set-shifting in the rat. Neurosci. 345:287-296. doi: 10.1016/j.neuroscience.2016.08.008.

Wassum KM, Ostlund SB, Maidment NT. (2012) Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. Biol Psychiatry. 71(10):846-54. doi: 10.1016/j.biopsych.2011.12.019.

Wise RA. (2005) Forebrain substrates of reward and motivation. J Comp Neurol. 493(1):115-21. doi: 10.1002/cne.20689.

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.

Recording and stimulation

In order to compare activity along the mediolateral axis of the dopaminergic nuclei, an array of 4 recording probes was constructed attached to 2 optical stimulation probes, with each serving the medial or lateral pair of recording probes (see Supp. Fig. 1 of Oberto, Matsumoto, et al., 2023). 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 main body of the headstage is composed of a resin (MadeSolid, Oakland, USA) molded in a 3D printer. The drives were adapted from the "lambda drive" of Headley, et al (2015). Two optic fibers were inserted in the headstage, which positioned the medial two at a depth of 3.9 mm above VTA, and the third and fourth at 3.7 and 3.5 mm above SNc, with the tips facing rostrally. The entire assembly is shielded from electrical noise with a fine mesh spot-glued with dental cement. This was connected to ground, serving as a miniature Faraday cage. Recordings 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.

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. The electrode assembly described above was implanted into the left VTA and SN. 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 either lit or off above the respective odor ports, again in pseudo-random sequence. Between the ports and array 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).

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. The headstage was then plugged in, and the mouse was placed into one compartment of the conditioning chamber. The room lights were dimmed and the recording system was turned on 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

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).

For tallies of runs of successive trials where the animals adhered to strategies other than the current rule, two methods were applied. For the strict "100% adherence" approach, we

counted occurrences of at least successive six trials adhering to the strategy. For the more permissive “>87%” 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). Trials while the mouse reached criterion performance for performing the rewarded strategy were not included in tallies of total number of trials here. Epochs with 14 or fewer (non-criterion) trials were excluded from this analysis. If more than one RP cell was recorded in an epoch, the epoch was only counted once for these tallies.

Overall, only 14.3% of trials were at criterion performance levels prior to a rule shift, and thus mice were training most of the time. 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 was ended. 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. The weight of the rewards earned was calculated, and the required amount of chow for the daily feeding was calculated and provided.

Data analyses

Single unit discrimination

Offline spike sorting was carried out with KiloSort (Pachitariu et al., 2016) followed by manual curation using Klusters (L. Hazan, http://neurosuite.sourceforge.net).

Neuron categories

To identify neurons as dopaminergic, we used the Stimulus-Associated spike Latency Test (SALT; Kvitsiani et al., 2013; Eshel et al., 2013). 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 analyses and behavioral correlates

Data were analyzed for [-500 ms, 500 ms] periods around three trial events. "Cue" is 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. Post-reward signal excitatory responses were noted when the firing rate in at least one bin after reward onset exceeded baseline activity by at least 50%. Baseline was the firing rate in the period [cue onset-2.0 s, cue onset-1.5 s]. Post-punishment signal inhibitory responses were noted when the firing rate in at least one bin was inferior to the baseline firing rate by at least 33% in the one second period after the punishment onset.

Statistical analyses

All tests were performed in MATLAB except for the chi square test which was performed on https://www.socscistatistics.com/tests/chisquare2/default2.aspx

Histology

Once stable recordings were no longer possible, marking lesions were made with 10 s of 30 µA cathodal current. After waiting for at least 90 min, mice were then killed with a lethal intraperitoneal injection of sodium pentobarbital, and perfused intra-ventricularly with phosphate buffered saline solution (PBS), followed by 10% phosphate buffered formalin. The brain was removed, post-fixed overnight, and placed in phosphate buffered 30% sucrose solution for 2-3 days. Frozen sections were cut at 80 µm, and permeabilized in 0.2% Triton in PBS for 1 h at room temperature. Antibody staining was performed to confirm localization of electrodes in dopaminergic nuclei. Sections were then treated with 3% bovine serum albumen (BSA) and 0.2% Triton in PBS for 1 h with gentle agitation at room temperature to block non-specific binding. Sections were then rinsed for 5 min in PBS at room temperature with gentle 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., 2022). Anatomical reconstruction with fluorescence microscopy revealed RP neuron recording within the ventral tegmental area at mediolateral positions ranging from 375 to 517 µm.

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