bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task

Iku Tsutsui-Kimura[1], Hideyuki Matsumoto[1,2], Naoshige Uchida[1] and Mitsuko Watabe-Uchida[1,3,*]

Affiliations:

[1]Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University, Cambridge, MA 02138, USA

[2]Department of Physiology, Osaka City University Graduate School of Medicine, Osaka, 545-8585, Japan

[3]Lead Contact

*Correspondence: mitsuko@mcb.harvard.edu (M.W.-U.)

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SUMMARY

Different regions of the striatum regulate different types of behavior. However, how dopamine signals differ across striatal regions and how dopamine regulates different behaviors remain unclear. Here, we compared dopamine axon activity in the ventral, dorsomedial, and dorsolateral striatum, while mice performed in a perceptual and value-based decision task. Surprisingly, dopamine axon activity was similar across all three areas. At a glance, the activity multiplexed different variables such as stimulus-associated values, confidence and reward feedback at different phases of the task. Our modeling demonstrates, however, that these modulations can be inclusively explained by moment-by-moment *changes* in the expected reward, i.e. the temporal difference error. A major difference between these areas was the overall activity level of reward responses: reward responses in dorsolateral striatum (DLS) were positively shifted, lacking inhibitory responses to negative prediction error. Tenets of habit and skill can be explained by this positively biased dopamine signal in DLS.

Keywords

dopamine, TD error, confidence, value, striatum, choice, feedback

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Flexibility in behavior relies critically on an animal's ability to alter behavior based on past experiences. In particular, the behavior of the animal is greatly shaped by the consequences of specific actions – whether a previous action led to positive or negative experiences. One of the fundamental questions in neuroscience is how animals learn from rewards and punishments.

A neurotransmitter, dopamine, is thought to be a key regulator of learning from rewards and punishments (Hart et al., 2014; Montague et al., 1996; Schultz et al., 1997). Neurons that release dopamine (hereafter, dopamine neurons) are located mainly in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc). These neurons send their axons to various regions including the striatum, neocortex, and amygdala (Menegas et al., 2015; Yetnikoff et al., 2014). The striatum, which receives the densest projection from VTA and SNc dopamine neurons, is thought to play particularly important roles in learning from rewards and punishments (Lloyd and Dayan, 2016; O'Doherty et al., 2004). However, what information dopamine neurons convey to the striatum, and how dopamine regulates behavior through its projections to the striatum remain elusive.

A large body of experimental and theoretical studies have suggested that dopamine neurons signal reward prediction errors (RPEs) – the discrepancy between actual and predicted rewards (Bayer and Glimcher, 2005; Cohen et al., 2012; Hart et al., 2014; Schultz et al., 1997). In particular, the activity of dopamine neurons resembles a specific type of prediction error, called temporal difference RPE (TD error) (Montague et al., 1996; Schultz et al., 1997; Sutton, 1988; Sutton and Barto, 1987). Although it was widely assumed that dopamine neurons broadcast homogeneous RPEs to a swath of dopamine-recipient areas, recent findings indicated that dopamine signals are more diverse than previously thought (Brown et al., 2011; Kim et al., 2015; Matsumoto and Hikosaka, 2009; Menegas et al., 2017, 2018; Parker et al., 2016). For one, recent studies have demonstrated that a transient ("phasic") activation of dopamine neurons occurs near the onset of a large movement (e.g. locomotion), regardless of whether these movements are immediately followed by a reward (Howe and Dombeck, 2016; da Silva et al., 2018). These phasic activations at movement onsets have been observed in the somatic spiking activity in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SNc (da Silva et al., 2018) as well as the axonal activity in the dorsal striatum (Howe and Dombeck, 2016). Another study showed that dopamine axons in the dorsomedial striatum (DMS) are activated when the animal makes a contralateral orienting movement in a decision-making task (Parker et al., 2016). Other studies have also found that dopamine axons in the posterior or ventromedial parts of the striatum are activated by aversive or threat-related stimuli (de Jong et al., 2019; Menegas et al., 2017). An emerging view is that dopamine neurons projecting to different parts of the striatum convey distinct signals and support different functions (Cox and Witten, 2019).

Previous studies have shown that different parts of the striatum control distinct types of reward-oriented behaviors (Dayan and Berridge, 2014; Graybiel, 2008; Malvaez and Wassum, 2018; Rangel et al., 2008). First, the ventral striatum (VS) has often been associated with Pavlovian behaviors, where the expectation of reward triggers relatively pre-programmed behaviors (approaching, consummatory behaviors etc.) (Dayan and Berridge, 2014). Psychological studies suggest that these behaviors are driven by stimulus-outcome associations (Kamin, 1969; Pearce and Hall, 1980; Rescorla and Wagner, 1972). Consistent with this idea, previous experiments have shown that dopamine in VS conveys canonical RPE signals (Menegas et al., 2017; Parker et al., 2016), and support learning of values associated with specific stimuli (Clark et al., 2012). In contrast, the dorsal part of the striatum has been linked to instrumental behaviors, where animals acquire an arbitrary action that leads to a reward (Montague et al., 1996; Suri and Schultz, 1999). Instrumental behaviors are further divided into two distinct types: goal-directed and habit (Dickinson and Weiskrantz, 1985). Goal-directed behaviors are "flexible" reward-oriented behaviors that are sensitive to a causal relationship ("contingency") between action and outcome, and can quickly adapt to changes in the value of the outcome (Balleine and Dickinson, 1998). After repetition of a goal-directed behavior, the behavior can become a habit which is characterized by insensitivity to changes in the outcome value (e.g. devaluation) (Balleine and O'Doherty, 2010). According to psychological theories, goal-directed and habitual behaviors are supported by distinct internal representations: action-outcome and stimulus-response associations, respectively (Balleine and O'Doherty, 2010). Lesion studies have indicated that goal-directed behaviors and habit are controlled by DMS and the dorsolateral striatum (DLS), respectively (Yin et al., 2004, 2005).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Instrumental behaviors are shaped by reward, and it is generally thought that dopamine is involved in their acquisition (Gerfen and Surmeier, 2011; Montague et al., 1996; Schultz et al., 1997). However, how dopamine is involved in distinct types of instrumental behaviors remain unknown. A prevailing view in the field is that habit is controlled by "model-free" reinforcement learning, while goal-directed behaviors are controlled by "model-based" mechanisms (Daw et al., 2005; Dolan and Dayan, 2013; Rangel et al., 2008). In this framework, habitual behaviors are driven by "cached" values associated with specific actions (action values) which animals learn through direct experiences via dopamine RPEs. In contrast, goal-directed behaviors are controlled by a "model-based" mechanism whereby action values are computed by mentally simulating which sequence of actions lead to which outcome using a relatively abstract representation (model) of the world. Model-based behaviors are more flexible compared to model-free behaviors because a model-based mental simulation may allow the animal to compute values in novel or changing circumstances. Although these ideas account for the relative inflexibility of habit over model-based, goal-directed behaviors, they do not necessarily explain the most fundamental property of habit, that is, its insensitivity to changes in outcome, as cached values can still be sensitive to RPEs when the actual outcome violates expectation, posing a fundamental limit in this framework (Dezfouli and Balleine, 2012; Miller et al., 2019). Furthermore, the idea that habits are supported by action value representations does not necessarily match with the long-held view of habit based on stimulus-response associations.

Until recently an implicit assumption across many studies was that dopamine neurons broadcast the same teaching signals throughout the striatum to support different kinds of learning (Rangel et al., 2008; Samejima and Doya, 2007). However, as mentioned before, more recent studies revealed different dopamine signals across striatal regions, raising the possibility that different striatal regions receive distinct teaching signals. In any case, few studies have directly examined the nature of RPE signals across striatal regions in instrumental behaviors, in particular, between DLS and other regions. As a result, it remains unclear whether different striatal regions receive distinct dopamine signals during instrumental behaviors. Are dopamine signals in particular areas dominated by movement-related signals? Are dopamine signals in these areas still consistent with RPEs or are they fundamentally distinct? How are they different? Characterizing

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dopamine signals in different regions is a critical step toward understanding how dopamine may regulate distinct types of behavior.

In the present study, we sought to characterize dopamine signals in different striatal regions (VS, DMS and DLS) during instrumental behaviors. We used a task involving both perceptual and value-based decisions in freely-moving mice – a task that is similar to those previously used to probe various important variables in the brain such as values, biases (Rorie et al., 2010; Wang et al., 2013), confidence (Hirokawa et al., 2019; Kepecs et al., 2008), belief states (Lak et al., 2017), and response vigor (Wang et al., 2013). In this task, the animal goes through various movements and mental processes – self-initiating a trial, collecting sensory evidence, integrating the sensory evidence with reward information, making a decision, initiating a choice movement, committing to an option and waiting for reward, receiving an outcome of reward or no reward, and adjusting internal representations for future performance using RPEs and confidence. Compared to Pavlovian tasks, which have been more commonly used to examine dopamine RPEs, the present task has various factors with which to contrast dopamine signals between different areas.

Contrary to our initial hypothesis, dopamine signals in all three areas showed similar dynamics, going up and down in a manner consistent with TD errors, reflecting moment-by-moment *changes* in the expected future reward (i.e. state values). Notably, although we observed correlates of accuracy and confidence in dopamine signals, consistent with previous studies (Engelhard et al., 2019; Lak et al., 2017), the appearance of these variables was timing- and trial type-specific. In stark contrast, our modeling demonstrate that these apparently diverse dopamine signals can be inclusively explained by a single variable – TD error, that is moment-by-moment changes in the expected reward in each trial. In addition, we found consistent differences between these areas. For instance, DMS dopamine signals were modulated by contralateral orienting movements, as reported previously (Parker et al., 2016). Furthermore, DLS dopamine signals, while following TD error dynamics, were overall more positive, compared to other regions. Based on these findings, we present novel models of how these distinct dopamine signals may give rise to distinct types of behavior such as flexible versus habitual behaviors.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

A perceptual decision-making task with reward amount manipulations

Mice were first trained in a perceptual decision-making task using olfactory stimuli (Figure 1) (Uchida and Mainen, 2003). To vary the difficulty of discrimination, we used two odorants mixed with different ratios (Figure 1A). Mice were required to initiate a trial by poking their nose into the central odor port, which triggered a delivery of an odor mixture. Mice were then required to move to the left or right water port depending on which odor was dominant in the presented mixture. Odor-water side (left or right) rule was held constant throughout training and recording in each animal. In order to minimize temporal overlaps between different trial events and underlying brain processes, we introduced a minimum time required to stay in the odor port (for 1 s before exiting the odor port) and in the water port (for 1 s) to receive a water reward.

After mice learned the task, the water amounts at the left and right water ports were manipulated (Lak et al., 2017; Rorie et al., 2010; Wang et al., 2013) in a probabilistic manner. In our task, one of the reward ports was associated with a big or medium size of water (BIG side) while another side was associated with a small or medium size of water (SMALL side) (Figure 1A). In a daily session, there were two blocks of trials, the first with equal-sized water and the second with different distributions of water sizes on the two sides (BIG versus SMALL side). The reward ports for BIG or SMALL conditions stayed unchanged within a session, and were randomly chosen for each session. In each reward port (BIG or SMALL side), which of the two reward sizes was delivered was randomly assigned in each trial. Note that the medium-sized reward is delivered with the probability of 0.5 for every correct choice at either side. This design was used to facilitate our ability to characterize RPE-related responses even after mice were well trained (Tian et al., 2016). First, the responses to the medium sized-reward allowed us to characterize how “reward expectation” affects dopamine reward responses because we can examine how different levels of expectation, associated with the BIG and SMALL side, affect dopamine responses to reward of the same (medium) amount. Conversely, for a given reward port, two

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sizes of reward allow us to characterize the effect of "actual reward" on dopamine responses, by comparing the responses when the actual reward was smaller versus larger than expected.

We first characterized the choice behavior by fitting a psychometric function (a logistic function). Compared to the block with equal-sized water, the psychometric curve was shifted laterally to the BIG side (Figure 1B, Figure S1). The fitted psychometric curves were laterally shifted whereas the slopes were not significantly different across blocks ($p=0.45$) (Figure 1B). We, therefore, quantified a choice bias as a lateral shift of the psychometric curve with a fixed slope in terms of the % mixture of odors for each mouse (Figure 1C) (Wang et al., 2013). All the mice exhibited a choice bias toward the BIG side (22/22 animals). Because a "correct" choice (i.e. whether a reward is delivered or not) was determined solely by the stimulus in this task, biasing their choices away from the 50/50 boundary inevitably lowers the choice accuracy (or equivalently, the probability of reward). For ambiguous stimuli, however, mice could go for a big reward, even sacrificing accuracy, in order to increase the long-term gain. Indeed, the observed biases approximated the optimal bias that maximizes total reward (1.016 ± 0.001 times reward compared to no bias, mean ± s.e.m, slightly less than the optimal bias that yields 1.022 times reward compared to no bias), rather than maximizing the accuracy (= reward probability, i.e. no bias) or solely minimizing the risk (the variance of reward amounts) (Figures 1D and 1E).

Previous studies have shown that animals shift their decision boundary even without reward amount manipulations in perceptual decision tasks (Lak et al., 2020a). These shifts occur on a trial-by-trial basis, following a win-stay strategy, choosing the same side when that side was associated with reward in the previous trial, particularly when the stimulus was more ambiguous (Lak et al., 2020a). In the current task design, however, the optimal bias is primarily determined by the sizes of reward (more specifically, which side delivered a big or small reward) which stays constant across trials within a session. To determine whether the animal adopted a short-time scale updating or a more stable bias, we next examined how receipt of reward affected the choice in the subsequent trials. To extract trial-by-trial updating, we compared the psychometric curves 1 trial before (n-1) and after (n+1) the current trials (n). This analysis was performed separately for the rewarded side in the current (n) trials. We found that choice biases before and after a specific reward location were not significantly different in any trial types (Figure 1F),

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

231 suggesting that trial-by-trial updating was minimum, contrary to a previous study (Lak et al.,
232 2020b). Instead, these results indicate that the mice adopted a relatively stable bias that lasts
233 longer than one trial.
234
235 Although we imposed a minimum time required to stay in the odor port, the mice showed
236 different reaction times (the duration between odor onset and odor port exit) across different trial
237 types (Figure 1G). First, reaction times were shorter when animals chose the BIG side compared
238 to the SMALL side in easy, but not difficult, trials. Second, reaction times were positively
239 correlated with the level of sensory evidence for choice (as determined by odor % for the choice)
240 when mice chose the BIG side. However, this modulation was not evident when mice chose the
241 SMALL side.
242
243 **Overall activity pattern of dopamine axons in the striatum**
244
245 To monitor the activity of dopamine neurons in a projection specific manner, we recorded the
246 dopamine axon activity in the striatum using a calcium indicator, GCaMP7f (Dana et al., 2019)
247 with fiber fluorometry (Kudo et al., 1992) (fiber photometry) (Figure 2). We targeted a wide
248 range of the striatum including the relatively dorsal part of VS, DMS and DLS (Figure 2B).
249 Calcium signals were monitored from mice both before and after introducing water amount
250 manipulations (n = 9, 7, 6 mice, for VS, DMS, DLS).
251
252 The main analysis was performed using the calcium signals obtained in the presence of water
253 amount manipulations. To isolate responses that are time-locked to specific task events but with
254 potentially overlapping temporal dynamics, we first fitted dopamine axon activity in each animal
255 with a linear regression model using multiple temporal kernels (Park et al., 2014) with Lasso
256 regularization with 10-fold cross validation (Figure 2). We used kernels that extract stereotypical
257 time courses of activity locked to four different events: odor onset (odor), odor port exit
258 (movement), water port entry (choice commitment or "choice" for short), and reward delivery
259 (water) (Figures 2C-2F).
260

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The constructed model captured modulations of dopamine axon activity time-locked to different events (Figure 2C). On average, the magnitude of the extracted odor-locked activity was modulated by odor cues. Dopamine axons were more excited by a pure odor associated with the BIG side than a pure odor associated with the SMALL side (Figures 2C and 2F). The movement-locked activity was stronger for a movement toward the contra-lateral, compared to the ipsi-lateral side, which was most evident in DMS (Parker et al., 2016) but much smaller in VS or DLS (Figure 2E, %Explained by movement). The choice-locked activity showed two types of modulations (Figure 2C). First, it exhibited an inhibition in error trials at the time of reward (i.e. when it has become clear that reward is not going to come). Second, dopamine activity showed a modulation around the time of water port entry, an excitation when the choice was correct, and an inhibition when the choice was incorrect, even before the mice received a feedback. These "choice commitment"-related signals will be further analyzed below. Finally, delivery of water caused a strong excitation which was modulated by the reward size (Figures 2C and 2F). Furthermore, the responses to medium-sized water was slightly but significantly smaller on the BIG side compared to the SMALL side (Figures 2C and 2F). The contribution of water-locked kernels was larger than other kernels except in DMS, where odor, movement and water kernels contributed similarly (Figures 2D and 2E).

In previous studies, RPE-related signals have typically been characterized by phasic responses to reward-predictive cues and a delivery or omission of reward. Overall, the above results demonstrate that observed populations contain the basic response characteristics of RPEs. First, dopamine axons were excited by reward-predicting odor cues, and the magnitude of the response was stronger for odors that instructed the animal to go to the side which is associated with a higher value (i.e. BIG side). Responses to water were modulated by reward amounts, and the water responses were suppressed by higher reward expectation. These characteristics were also confirmed by using the actual responses, instead of obtained kernel models (Figures 2F and 2G). Finally, in error trials, dopamine axons were inhibited when the time passed beyond the expected time of reward, as the negative outcome becomes certain (Figure 2C). Next, we will investigate each striatal area in more detail.

Shifted representation of TD error in dopamine axon activity across the striatum

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Although excitation to unpredicted reward is one of the signatures of dopamine RPE, recent studies found that the dopamine axon response to water is small or undetectable in some part of the dorsal striatum (Howe and Dombeck, 2016; Parker et al., 2016; da Silva et al., 2018). Therefore, the above observation that all three areas (VS, DMS, and DLS) exhibited modulation by reward may appear at odds with previous studies.

We noticed greatly diminished water responses when the reward amount was not manipulated, that is, when dopamine axon signals were monitored during training sessions before introducing the reward amount manipulations (Figure 3). In these sessions, dopamine axons in some animals did not show significant excitation to water rewards (Figures 3A and 3D). This "lack" of reward response was found in DMS, consistent with previous studies (Parker et al., 2016), but not in VS or DLS (Figure 3G). Surprisingly, however, DMS dopamine axons in the same animals showed clear excitation when reward amount manipulations were introduced, particularly strongly responding to a big reward (Figures 3B and 3E). Indeed, the response patterns were qualitatively similar across different striatal areas (Figure 4); the reward responses in all the areas were modulated by reward size and expectation, although the whole responses seem to be shifted higher in DLS, and lower in DMS (Figures 4A and 4B). These results indicate that a stochastic nature of reward delivery in our task enhanced or "rescued" reward responses in dopamine axons in DMS.

The above results emphasized the overall similarity of reward responses across areas, but some important differences were also observed. Most notably, although a delivery of a small reward caused an inhibition of dopamine axons below baseline in VS and DMS, the activity remained non-negative in DLS. The overall responses tended to be higher in DLS.

In order to understand the diversity of dopamine responses to reward, we examined modulation of dopamine activity by different parameters (Figure 4D). First, the effect of the amount of "actual" reward was quantified by comparing responses to different amounts of water for a given cue (i.e. the same expectation). The reward responses in all areas were modulated by reward amounts, with a slightly higher modulation by water amounts in VS (Figure 4D Water big-

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

medium, Water medium-small). Next, the effect of expectation was quantified by comparing the responses to the same amounts of water with prediction of different amounts. Effects of reward size prediction were not significantly different across areas, although VS showed slightly less modulation with more variability (Figure 4D, prediction SMALL-BIG).

Next, we sought to characterize these differences between areas in simpler terms by fitting response curves (response functions). Previous studies that quantified responses of dopamine neurons to varied amounts of reward under different levels of expectation indicated that their reward responses can be approximated by a common function, with different levels of expectation just shifting the resulting curves up and down while preserving the shape (Eshel et al., 2016). We, therefore, fitted dopamine axon responses with a common response function (a power or linear function) for each expectation level (i.e. separately for BIG and SMALL) while fixing the shape of the function (i.e. the exponent of the power function or the slope of the linear function were fixed, respectively) (Figure 4C, Figure S2A). The obtained response functions for the three areas recapitulated the main difference between VS, DMS and DLS, as discussed above. For one, the response curves of DLS are shifted overall upward. This can be characterized by estimating the amount of water that does not elicit a change in dopamine responses from baseline firing ("zero-crossing point" or reversal point). The zero-crossing points, obtained from the fitted curves, were significantly lower in DLS (Figures 4C and 4D). The results were similar regardless of whether the response function was a power (power function $\alpha < 1$) or a linear function ($\alpha = 1$) (Figure S2B). Similar results were obtained using the aforementioned kernel models in place of the actual activity (Figure S2D).

Since the recording locations varied across animals, we next examined the relationship between recording locations and the zero-crossing points (Figures 4E and 4F). The zero-crossing points varied both along the medial-lateral and the dorsal-ventral axes (linear regression coefficient; β = -44 [zero-crossing point water amounts/mm], $p = 0.008$ for medial-lateral axis; β = -52, $p = 0.011$ for the dorsal-ventral axis). Examination of each animal confirmed that DMS showed higher zero-crossing points (upper-left in Figure 4E left) whereas DLS showed lower zero-crossing points (upper-right cluster in Figure 4E right).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We next examined whether the difference in zero-crossing points manifested specifically during reward responses or whether it might be explained by recording artifacts; upward and downward shifts in the response function can be caused by a difference in baseline activity before trial start (odor onset), and/or lingering activity of pre-reward activity owing to the relatively slow dynamics of the calcium signals (a combination of calcium concentration and the indicator). To examine these possibilities, the same analysis was performed after subtracting the pre-reward signals (Figure S2C). We observed similar or even bigger differences in zero-crossing points ($p=2.2\times10^{-5}$, analysis of variance [ANOVA]). These results indicate that the elevated or decreased responses, characterized by different zero-crossing points, was not due to a difference in "baseline" but was related to the difference that manifests specifically in responses to reward.

Considerably small zero-crossing points in dopamine axons in DLS were not due to a poor sensitivity to reward amounts nor a poor modulation by expected reward (Figure 4D). Different zero-crossing points, i.e. shifts of the boundary between excitation and inhibition at reward, suggest biased representation of TD error in dopamine axons across the striatum. In TD error models, difference in zero-crossing points may affect not only water responses but also responses to other events. Thus, the small zero-crossing points in dopamine axons in DLS should yield almost no inhibition following an event that is worse than predicted. To test this possibility, we examined responses to events with lower value than predicted (Figure 5): small water (Figures 5A-5C), water omission caused by choice error (Figures 5D-5F), and a cue that was associated with no outcome (Figures 5G-5I). Consistent with our interpretation of small zero-crossing points, dopamine axons in DLS did not show inhibition in response to outcomes that were worse than predicted while being informative about water amounts.

Taken together, these results demonstrate that dopamine reward responses in all three areas exhibited characteristics of RPEs. However, relative to canonical responses in VS, the responses were shifted more positively in the DLS and more negatively in the DMS.

TD error dynamics in signaling perceptual uncertainty and cue-associated value

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The analyses presented so far mainly focused on phasic dopamine responses time-locked to cues and reward. However, dopamine axon activity also exhibited richer dynamics between these events, which need to be explained. For instance, the signals diverged between correct and error trials even before the actual outcome was revealed (a reward delivery versus a lack thereof) (Figure 2C Choice). This difference between correct and error trials, which is dependent on the strength of sensory evidence (or stimulus discriminability), was used to study how neuronal responses are shaped by "confidence". Confidence is defined as the observer's posterior probability that their decision is correct given their subjective evidence and their choice ($P(reward|stimulus, choice)$) (Hangya et al., 2016). A decision model allows the experimenter to link stimulus discriminability to subjective evidence (Hangya et al., 2016). A given model and task structure makes specific predictions on the shape of three key signatures relating stimulus discriminability, choice and confidence. The predictions can vary depending on task design (Adler and Ma, 2018; Rausch and Zehetleitner, 2019), but the structure of our task follows the original predictions (Hangya et al., 2016). Additionally, in our task, the mice combined the information about reward size with the strength of sensory evidence to select an action (confidence, or uncertainty) (Figure 1). The previous analyses did not address how these different types of information affect dopamine activity over time. We next sought to examine the time course of dopamine axon activity in greater detail, and to determine whether a simple model could explain these dynamics.

Our task design included two delay periods, imposed before choice movement and water delivery, to improve our ability to separate neuronal activity associated with different processes (Figure 1A). The presence of stationary moments before and after the actual choice movement allows us to separate time windows before and after the animal's commitment to a certain option. We examined how the activity of dopamine neurons changed before choice movement and after the choice commitment (Figure 6).

We first examined dopamine axon activity after water port entry (0-1 s after water port entry). In this period, the animals have committed to a choice and are waiting for the outcome to be revealed. Responses following different odor cues were plotted separately for trials in which the animal chose the BIG or SMALL side. The vevaiometric curve (a plot of responses against

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sensory evidence) followed the expected 'X-pattern' with a modulation by reward size (Hirokawa et al., 2019), which matches the expected value for these trial types, or the size of reward multiplied by the probability of receiving a reward, given the presented stimulus and choice (Figure 6C). The latter has been interpreted as the decision confidence, $P(reward|stimulus, choice)$ (Lak et al., 2017, 2020b). The crossing point of the two lines forming an "X" is shifted to the left in our data because of the difference in the reward size (Figure 6C).

When this analysis was applied to the time period before choice movement (0-1 s before odor port exit), the pattern was not as clear; the activity was monotonically modulated by the strength of sensory evidence (%Odor BIG) only for the BIG choice trials, but not for the SMALL choice trials (Figure 6B). This result is contrary to a previous study that suggested that the dopamine activity reflecting confidence develops even before a choice is made (Lak et al., 2017). We note, however, that the previous study only examined the BIG choice trials, and the results were shown by "folding" the x-axis, that is, by plotting the activity as a function of the stimulus contrast (which would correspond to |%Odor BIG – 50| in our task), with the result matching the so-called "folded X-pattern". We would have gotten the same result, had we plotted our results in the same manner excluding the SMALL choice trials. Our results, however, indicate that a full representation of "confidence" only becomes clear after a choice commitment, leaving open the question what the pre-choice dopamine activity really represents.

The aforementioned analyses, using either the kernel regression or actual activity showed that cue responses were modulated by whether the cue instructed a choice toward the BIG or SMALL side (Figures 2C and 2F). These results indicate that the information about stimulus-associated values (BIG versus SMALL) affected dopamine neurons earlier than the strength of sensory evidence (or confidence). We next examined the time course of how these two variables affected dopamine axon activity more closely. We computed the dopamine axon activity between trials when a pure odor instructed to go to the BIG versus SMALL side. Consistent with the above result, the difference was evident during the cue period, and then gradually decreased after choice movement (Figure 6D). We performed a similar analysis, contrasting between easy and difficult trials (i.e. the strength of sensory evidence). We computed the difference between

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dopamine axon activity in trials when the animal chose the SMALL side after the strongest versus weaker stimulus evidence (a pure odor that instruct to choose the SMALL side versus an odor mixture that instruct to choose the BIG side). In stark contrast to the modulation by the stimulus-associated value (BIG versus SMALL), the modulation by the strength of stimulus evidence in SMALL trials fully developed only after a choice commitment (i.e. water port entry) (Figure 6E). Across striatal regions, the magnitude and the dynamics of modulation due to stimulus-associated values and the strength of sensory evidence were similar (Figures 6F and 6G), although we noticed that dopamine axons in DMS showed slightly higher correlation with sensory evidence before choice (Figure S3).

As discussed above, a neural correlate of "confidence" appears at a specific time point (after choice commitment and before reward delivery) or in a specific trial type (when an animal would choose BIG side) before choice. We, therefore, next examined whether a simple model can account for dopamine axon activity more inclusively (Figure 7). To examine how the value and RPE may change within a trial, we employed a Monte-Carlo approach to simulate animal's choices assuming that the animal has already learned the task. We used a Monte-Carlo method to obtain the ground truth landscape of the state values over different task states, without assuming a specific learning algorithm.

The variability and errors in choice in psychophysical performance are thought to originate in the variability in the process of estimating sensory inputs (perceptual noise) or in the process of selecting an action (decision noise). We first considered a simple case where the model contains only perceptual noise (Green and Swets, 1966). In this model, an internal estimate of the stimulus or a "subjective odor" was obtained by adding Gaussian noise to the presented odor stimulus on a trial-by-trial basis (Figures 7B left). In each trial, the subject chooses deterministically the better option (Figure 7C left) based on the subjective odor and the reward amount associated with each choice (Figure 7B right). The model had different "states" considering N subjective odors (N = 60 and 4 were used and yielded similar results), the available options (left versus right), and a sequence of task events (detection of odor, recognition of odor identity, choice movement, water port entry [choice commitment], Water/No-water feedback, inter-trial interval [ITI]) (Figure 7A). The number of available choices is two after

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

478 detecting an odor but reduced to 1 (no choice) after water port entry. In each trial, the model
479 receives one of the four odor mixtures, makes a choice, and obtains feedback (rewarded or not).
480 After simulating trials, the state value for each state was obtained as the weighted sum of
481 expected values of the next states, which was computed by multiplying expected values of the
482 next states with probability of transitioning into the corresponding state. After learning, the state
483 value in each state approximates the expected value of future reward, sum of the amount of
484 reward multiplied by probability of the reward (for simplicity, we assumed no temporal
485 discounting of value within a trial). After obtaining state values for each state, state values for
486 each odor ("objective" odor presented by experimenters) was calculated as the weighted sum of
487 state values over subjective odors. After obtaining state values at each state, we then computed
488 TD errors using a standard definition of TD error which is the difference between the state values
489 at consecutive time points plus received rewards at each time step (Sutton and Barto, 1987).
490
491 We first simulated the dynamics of state values and TD errors when the model made a correct
492 choice in easy trials, choosing either the BIG or SMALL side (Figure 7F bottom, blue versus
493 red). As expected, the state values for different subjective odors diverged as soon as an odor
494 identity was recognized, and the differences between values stayed constant as the model
495 received no further additional information before acquisition of water. TD errors, which are the
496 derivative of state values, exhibited a transient increase after odor presentation, and then returned
497 to their baseline levels (near zero), remaining there until the model received a reward. Next, we
498 examined how the strength of sensory evidence affected the dynamics of value and TD errors
499 (Figures 7F and 7J). Notably, after choice commitment, TD error did not exhibit the additional
500 modulation by the strength of sensory evidence, or a correlate of confidence (Figures 7F right
501 and 7J right), contrary to our data (Figures 7E and 7I right). Thus, this simple model failed to
502 explain aspects of dopamine axon signals that we observed in the data.
503
504 In the first model, we assumed that the model picks the best option given the available
505 information in every trial (Figure 7C). In this deterministic model, all of the errors in choice are
506 attributed to perceptual noise. We next considered a model that included decision noise in
507 addition to the perceptual noise (Figure 7D). Here decision noise refers to some stochasticity in
508 the action selection process, and may arise from errors in an action selection mechanism or

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exploration of different options, and can be modeled using different methods or rationale behind it. Here we present results based on a "softmax" decision rule, in which a decision variable (in this case, the difference in the ratio of the expected values at the two options) was transformed into the probability of choosing a given option using a sigmoidal function (e.g. Boltzmann distribution) (Sutton and Barto, 2011). We also tested other stochastic decision rules such as Herrnstein's matching law (Herrnstein, 1961) or ε-greedy exploration (randomly selecting an action in a certain fraction [ε] of trials) (Sutton and Barto, 2011) (Figures S4A-S4C).

Interestingly, just by adding some stochasticity in action selection, various peculiar features of dopamine axon signals described above were suddenly explained (Figures 7G and 7K). Note that the main free parameters of the above models are the width of the Gaussian noise, which determines the "slope" of the psychometric curve, and was chosen based merely on the behavioral performance, but not the neural data. When the model chose the BIG side, state value at odor presentation was roughly monotonically modulated by the strength of sensory evidence similar to the above (Figure 7G top left). When the model chose the SMALL side, however, the relationship between the strength of sensory evidence and value was more compromised (Figure 7G middle left). As a result, TD error did not show monotonic relationship with sensory evidence before choice (Figures 7G middle right and 7K left), similar to actual dopamine axons responses (Figures 7E middle and 7I left), which was reminiscent of reaction time pattern (Figure 7H). On the other hand, once a choice was committed, the model exhibited interesting dynamics very different from the above deterministic model. After choice commitment, expected value was monotonically modulated by the strength of sensory evidence both for the choice to the BIG and SMALL sides (Figure 7G top and middle left, After). Further, because of the introduced stochasticity in action selection, the model sometimes chose a suboptimal option, resulting in a drop in the state value. This, in turn, caused TD error to exhibit an "inhibitory dip" once the model "lost" a better option (Figure 7G right), similar to the actual data (Figures 7E and 7I). This effect was strong particularly when the subjective odor instructed the BIG side but the model ended up choosing the SMALL side. For a similar reason, TD error showed a slight excitation when the model chose a better option (i.e. lost a worse option). The observed features in TD dynamics were not dependent on exact choice strategy: softmax, matching, and ε-greedy, all produced similar results (Figures S4B and S4C). This is because, with any strategy, after

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

commitment of choice, the model loses another option with a different value, which results in a change in state value. These results are in stark contrast to the first model in which all the choice errors were attributed to perceptual noise.

In summary, we found that a standard TD error, computing the moment-by-moment changes in state value (or, the expected future reward), can capture various aspects of dynamics in dopamine axon activity observed in the data, including the changes that occur before and after choice commitment, and the detailed pattern of cue-evoked responses. These results were obtained as long as we introduced some stochasticity in action selection (decision noise), regardless of how we did it. The state value dynamically changes during the performance of the task because the expected value changes according to an odor cue (i.e. strength of sensory evidence and stimulus-associated values) and the changes in potential choice options. A drop of the state value and TD error at the time of choice commitment occurs merely because the state value drops when the model chose an option that was more likely to be an error. Further, a correlate of "confidence" appears after committing a choice, merely because at that point (and *only* at that point), the state value becomes equivalent to the reward size multiplied with the confidence, i.e. the probability of reward given the stimulus and the choice. This means that, as long as the animal has appropriate representations of states, a representation of "confidence" can be acquired through a simple associative process or model-free reinforcement learning without assuming other cognitive abilities such as belief states or self-monitoring (meta-cognition). In total, not only the phasic responses but also some of the previously unexplained dynamic changes can be explained by TD errors computed over the state value, provided that the model contains some stochasticity in action selection in addition to perceptual noise. Similar dynamics across striatal areas (Figure 6) further support the idea that dopamine axon activity follows TD error of state values in spite of the aforementioned diversity in dopamine signals.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

In the present study, we monitored dopamine axon activity in three regions of the striatum (VS, DMS and DLS) while mice performed instrumental behaviors involving perceptual and value-based decisions. In addition to phasic responses associated with reward-predictive cues and reward, we also analyzed more detailed temporal dynamics of the activity within a trial. Contrary to the current emphases on diversity or multiplexing in dopamine signals (and therefore, to our surprise), we found that dopamine axon activity in all of the three areas exhibited dynamics that can be explained by the TD error which calculates moment-by-moment "changes" in the expected future reward (i.e. state value). Interestingly, however, our results showed consistent differences between regions. First, as reported previously (Parker et al., 2016), during choice movements, contra-lateral orienting movements caused a transient activation in the DMS. This response was negligible in VS and DLS, however. Second, although dopamine axon signals exhibited temporal dynamics that are predicted by TD errors, reward responses were generally elevated in DLS. As a consequence, dopamine axon signals in DLS did not exhibit a clear inhibitory response ("dopamine dip") even when the actual reward was smaller than expected, or even when the animal did not receive a reward, despite our observations that dopamine axons in VS and DMS exhibited clear inhibitory responses in these conditions. Overall, the activity during the reward period was biased toward positive responses in the DLS, compared to other areas. Activation of dopamine neurons both in VTA and SNc are known to reinforce preceding behaviors (Ilango et al., 2014; Keiflin et al., 2019; Lee et al., 2020; Saunders et al., 2018). The differences in dopamine axon signals that we observed in instrumental behaviors can provide specific constraints on the behaviors learned through dopamine-mediated reinforcement in these striatal regions.

Diversity in representation of TD errors

Accumulating evidence indicates that dopamine neurons are diverse in various aspects such as anatomy, physiological properties, and activity (Engelhard et al., 2019; Farassat et al., 2019; Howe and Dombeck, 2016; Kim et al., 2015; Lammel et al., 2008; Matsumoto and Hikosaka, 2009; Menegas et al., 2015, 2017, 2018; Parker et al., 2016; da Silva et al., 2018; Watabe-Uchida

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and Uchida, 2018). Our study is one of the first to examine dopamine signals in three different regions of the striatum during an instrumental behavior involving perceptual and value-based decisions. We found that dopamine axon activity in the striatum follows TD error dynamics in our choice paradigm. At the same time, we found that the response function for water delivery in dopamine axons in different striatal areas showed different zero-crossing points, the boundary between excitatory and inhibitory responses (Figure 4). The results suggested that dopamine axons in DMS use a higher boundary (requiring larger amounts of reward to excite), and dopamine axons in DLS use a lower boundary (requiring smaller amounts of reward to excite). In other words, dopamine signals in DMS use a strict criterium to be excited, whereas dopamine signals in DLS tend to be more excited with smaller rewards.

A recent study (Dabney et al., 2020) proposed that the diversity in dopamine responses potentially give rise to a population code for a reward distribution (distributional reinforcement learning). In this theory, there are optimistic and pessimistic dopamine neurons. Optimistic dopamine neurons emphasize positive over negative RPEs, and as a consequence, their corresponding value predictors are biased to predict a higher value in a reward distribution, or vice versa. The distributional reinforcement learning, as formulated in Dabney et al. (Dabney et al., 2020), predicts that optimistic and pessimistic dopamine neurons should have zero-crossing points shifted toward larger and smaller rewards, respectively. In this sense, our observation that DLS dopamine signals have smaller zero-crossing points resembles pessimistic dopamine neurons in distributional reinforcement learning, although the previous study found both optimistic and pessimistic dopamine neurons in the VTA, which does not necessarily project to the DLS. Whether the present result is related to distributional reinforcement learning requires more specific tests such as dopamine neurons' sensitivity to positive versus negative RPEs (Dabney et al., 2020). It will be interesting to characterize these response properties in a projection-specific manner.

Higher criteria in DMS may partly explain the observation that some dopamine neurons do not show a clear excitation by reward, such as in the case of our recording without reward amount modulations (Figure 3). Our results suggest that whether dopamine neurons respond to reward likely depends critically on task structures and training history. It will be important to further

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

examine in what conditions these dopamine neurons lose responses to water, or whether there are dopamine neurons which do not respond to reward in any circumstances. In contrast to DMS, we observed reliable excitation to water reward in dopamine axons in DLS. Thus, the previous observation that some dopamine neurons in the substantia nigra show small or no excitation to reward (da Silva et al., 2018) may mainly come from DMS-projecting dopamine neurons or another subpopulation of dopamine neurons that project to the tail of the striatum (TS) (Menegas et al., 2018), but not DLS. The distinction is important because smaller dopamine responses to reward have been often linked to skill or habit with value-free mechanism (Miller et al., 2019). In contrast, we found that dopamine axons in DLS show strong modulation by reward amounts and prediction, and its dynamics resemble TD errors. Our observation suggests that the lack of reward omission responses and excitation by even small rewards is a key for the function of dopamine in DLS.

Positively biased reinforcement signals in DLS dopamine

It has long been observed that the activity of many dopamine neurons exhibits a phasic inhibition when an expected reward was omitted or when the reward received was smaller than expected (Hart et al., 2014; Schultz et al., 1997). This inhibitory response to negative RPEs is one of the hallmarks of dopamine RPE signals. Our results that DLS dopamine signals largely lack these dopamine dips (Figure 4 and Figure 5) has profound implications on what types of behaviors are learned through DLS dopamine signals as well as what computational principles underlie reinforcement learning in DLS.

Dopamine "dips" are thought to act as aversive stimuli and/or can facilitate extinction of previously learned behaviors (weakening) (Chang et al., 2018; Montague et al., 1996; Schultz et al., 1997). The lack of dopamine dip in DLS may lead to the animal's reduced sensitivity to worse-than-expected outcome (i.e. negative prediction error). This characteristic resembles the activity of dopamine axons in TS, posterior to DLS, which signals potential threat and also lacks inhibitory responses to an omission of a predicted threat (Menegas et al., 2017, 2018). We proposed that the lack of inhibitory omission signals (and so lack of weakening signals) would

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be critical to maintain threat prediction even if an actual threat is sometimes omitted. Similarly, the lack of weakening signals in DLS may help keep the learned actions from being erased even if the outcome is sometimes worse than predicted or even omitted. This idea is in line with the previous observations that DLS plays an important role in habitual behaviors (Yin et al., 2004). The uniquely modified TD error signal in DLS (i.e. a reduced inhibitory response during the reward period) may explain a predominant role of DLS in controlling habitual behaviors.

What is learned in the DLS? "The law of exercise" and learning sequences.

A deeper understanding of the nature of reinforcement signals can constrain the search for computational principles and provide critical insight into what is actually learned by the system. Here we speculate on these questions in the light of reinforcement learning theories and anatomy.

Thorndike (Thorndike, 1932) proposed three principles for instrumental learning – the law of effect, the law of readiness, and the law of exercise. The law of effect emphasizes the role of outcome of behaviors: behaviors that led to good outcomes become more likely to occur – the idea that is a foundation of value-based reinforcement learning. In contrast, the law of exercise emphasizes the number of times a particular action was taken. There has been an increasing appreciation of the law of exercise because repetition or overtraining is the hallmark of habits and skills (Hikosaka et al., 1995; Matsuzaka et al., 2007; Miller et al., 2019; Morris and Cushman, 2019; Ölveczky, 2011; Robbins and Costa, 2017; Smith and Graybiel, 2016). Here we propose that dopamine signals in DLS provide an ideal neural substrate of learning with an emphasis on the law of exercise. A positively biased TD error signals ensures that an "OK" action will be positively reinforced, in a manner that depends on the number of times that the same behavior was repeated as far as it is accompanied by a small reward (i.e. with "OK" signals). This property may explain why the formation of habit (and skills) normally requires overtraining (i.e. repeating a certain behavior many times).

The observation that DLS dopamine signals lack inhibitory responses raises the question what is actually learned by the system. Learning of values depends on the balance between positive and

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

negative prediction errors: the learned value converges to the point at which positive and negative prediction errors form an equilibrium. If a reinforcement signal lacks negative prediction errors, this learning would no longer work as it was originally conceptualized. In reinforcement learning theories, an alternative approach is policy-based reinforcement learning. We propose that policy learning may be a better way to conceptualize the function of the DLS. In reinforcement learning, a policy is a set of rules that map an action to a state, and has direct relevance to stimulus-response associations that are proposed to underlie habit because the relationship between stimulus (state) and response (action) can be more directly encoded using a policy. According to Sutton and Barto (Sutton and Barto, 2018), policy learning can be done by learning what is called "preference", $h(s, a)$, which defines the likelihood of a certain action, a, in a given state, s. In a given state, an action is selected based on preference through a winner-take-all mechanism either deterministically (e.g. by selecting the action with the maximum preference) or stochastically (e.g. through a softmax action selection). One way to conceptualize preference is to see it as a generalized version of value, which has less constraints than value (the idea of "value" may imply many properties that it should follow, e.g. the value should be zero for no outcome). Alternatively, $h(s, a)$ can directly encode the probability of an action.

It is also important to consider what are "states" for learning in DLS. Importantly, the main inputs to DLS come from the motor cortex, somatosensory cortex, and other subcortical areas such as intralaminar nuclei in thalamus (Hunnicutt et al., 2016). Thus, the inputs to DLS may not be dominated by the sensory information representing the external world, as often conceptualized in reinforcement learning. Instead, DLS is well-positioned to receiving inputs representing motor commands (the current "motor states") or somatosensory information (the current "bodily states" consisting of proprioception, sense of touch etc.). In other words, DLS may compute their output by monitoring the current motor and bodily states. Dopamine in DLS can thus be conceptualized as a reinforcement signal that strengthens the connection between the current motor/bodily state and the next motor output. This mechanism, when chained, can produce a sequence of movements as long as the same motor/bodily state is revisited or reproduced, which may not occur easily at the beginning but can occur after repeated training. As such, DLS may regulate "how" to perform a sequence of well-trained movements smoothly and automatically. The key properties of habits and skills such as stereotypy, automaticity and

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

721 the requirement of overtraining, can be explained by this model. In this model, the learning of
722 habits and skills are a natural consequence of reinforcement learning using a specialized
723 reinforcement signals (positively shifted response to outcomes) and the unique anatomical
724 property (the specialized input suitable for chaining actions) of the DLS. Future experiments
725 using tasks involving sequence of actions (Hikosaka et al., 1995; Ölveczky, 2011) can test this
726 idea.
727
728 **Potential mechanisms underlying diverse TD error signals**
729
730 We found that, across the striatum, dopamine signals overall resemble TD errors, with positive
731 or negative biases in a subregion-specific manner (Figure 4). A potential mechanism to generate
732 such a diversity is by optimistic and pessimistic expectations, as proposed in distributional
733 reinforcement learning (Dabney et al., 2020). Alternatively, DLS-projecting dopamine neurons
734 may add "success premium" at each feedback. Signals of success feedback were observed in
735 multiple cortical areas (Chen et al., 2017; Sajad et al., 2019; Stuphorn et al., 2000), which is
736 often more sustained than phasic dopamine responses. Interestingly, we noticed that responses to
737 water in dopamine axons in DLS are more sustained than dopamine axons in other areas (Figure
738 4A). DLS-projecting dopamine neurons potentially receive and integrate those success feedback
739 signals with reward value, shifting the teaching signals more positively.
740
741 Mechanistically, biases in dopamine signals may stem from a difference in the excitation-
742 inhibition balance at the circuit level. In addition to dopamine neurons, there are multiple brain
743 areas where activity of some neurons resembles RPE (Li et al., 2019; Matsumoto and Hikosaka,
744 2007; Oyama et al., 2010; Tian et al., 2016). Among these, presynaptic neurons in multiple brain
745 areas directly convey a partial prediction error to dopamine neurons (Tian et al., 2016). On the
746 other hand, rostromedial tegmental area (RMTg) exhibits a flipped version of RPE (the sign is
747 opposite to dopamine neurons), and its inhibitory neurons directly project to dopamine neurons
748 in a topographic manner (Hong et al., 2011; Jhou et al., 2009a, 2009b; Li et al., 2019; Tian et al.,
749 2016). Hence, each dopamine neuron may receive a different ratio of excitatory and inhibitory
750 inputs of RPE. It would be interesting if DLS-projecting dopamine neurons receive less
751 inhibitory RPE, and DMS-projecting dopamine neurons receive more, so that RPE signals are

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pushed up or down, whereas the information is still almost intact. In addition to anatomical reasons, DLS-projecting dopamine neurons show higher burstiness in intact animals (Farassat et al., 2019) and higher excitability *in vitro* (Evans et al., 2017; Lerner et al., 2015). These multiple reasons may explain why DLS-projecting dopamine neurons do not show inhibitory responses to negative prediction errors. It will be fascinating if we could connect all these levels of studies into functional meaning in the future.

Future directions to understand the meaning of diversity of dopamine signals

Recent studies reported that dopamine neurons are modulated by various parameters (Engelhard et al., 2019; Watabe-Uchida and Uchida, 2018). Here, we found that TD error dynamics can inclusively explain two seemingly separate decision variables, namely, stimulus-associated value and choice accuracy when animal's choice strategy is not deterministic (i.e. there is decision noise) (Figure 6). At a glance, dopamine activity patterns may appear to be signaling two distinct variables at different timings, but both are inclusively explained by a single quantity (TD error) in one framework (Figure 7). These results underscore the importance of considering moment-by-moment dynamics, and underlying computation. Taken together, our results showed that dopamine axon signals in the striatum approximate TD error dynamics. We propose that dopamine in different striatal areas conveys TD errors in a biased manner. One compelling idea is that the lack of negative teaching signals in DLS plays a role in skill/habit, although further examination is needed to establish its functions. Although we designed the task to minimize effects of movement itself on results, accumulating studies suggested close relationship between dopamine signaling and movement (Howe and Dombeck, 2016; da Silva et al., 2018). It is important to test these other parameters in the future in order to understand the meaning of the diversity of dopamine neurons and organization of dopamine-striatum systems.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EXPERIMENTAL PROCEDURES

Animals

17 dopamine transporter (DAT)-cre (B6.SJL-Slc6a3tm1.1(cre)Bkmn/J, Jackson Laboratory; RRID:IMSR JAX:006660) (Bäckman et al., 2006) heterozygous mice, and 5 DAT-Cre;Ai14 (Rosa-CAG-LSL-tdTomato, Jackson Laboratory; RRID:IMSR JAX:007914) (Madisen et al., 2010) double heterozygous mice, male and female, were used for recording signals from dopamine axons. All mice were backcrossed with C57BL/6J (Jackson Laboratory). Animals were housed on a 12 hour dark/12 hour light cycle (dark from 07:00 to 19:00) and performed a task at the same time each day. All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Animal Care and Use Committee.

Surgical Procedures

All surgeries were performed under aseptic conditions with animals anesthetized with isoflurane (1–2% at 0.5–1.0 l/min). Analgesia was administered pre (buprenorphine, 0.1 mg/kg, I.P) and postoperatively (ketoprofen, 5 mg/kg, I.P). To express GCaMP7f (Dana et al., 2019) specifically in dopamine neurons, we unilaterally injected 300 nl of mixed virus solution; AAV5-CAG-FLEX-GCaMP7f (1×10^{12} particles/ml, UNC Vector Core, NC) and AAV5-CAG-FLEX-tdTomato (2×10^{13} particles/ml, UNC Vector Core, NC) into both the VTA and SNc (600 nl total) in the DAT-cre mice. Only AAV5-CAG-FLEX-GCaMP7f (300 nl total) was used for DAT;Ai14 double transgenic mice. Virus injection lasted around 20 minutes, and then the injection pipette was slowly removed over the course of several minutes to prevent damage to the tissue. We also implanted optic fibers (400 μm diameter, Doric Lenses, Canada) into the VS, DMS, or DLS (1 fiber per mouse). To do this, we first slowly lowered optical fibers into the striatum. Once fibers were lowered, we first attached them to the skull with UV-curing epoxy (NOA81, Thorlabs, NJ), and then a layer of rapid-curing epoxy to attach the fiber cannulas even more firmly to the underlying glue. After waiting 15 minutes for this to dry, we applied a black dental adhesive (Ortho-Jet, Lang Dental, IL). We used magnetic fiber cannulas (Doric Lesnses, MFC_400/430) and the corresponding patch cords to allow for recordings in freely moving

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

812 animals. After waiting 15 minutes for the dental adhesive to dry, the surgery was complete. We
813 used the following coordinates to target our injections and implants.
814
815 - (VTA) Bregma: -3.0 mm, Lateral: 0.6 mm, Depth: between 4.5 mm and 4.3 mm
816 - (SNc) Bregma: -3.0 mm, Lateral: 1.6 mm, Depth: between 4.3 mm and 4.1 mm
817 - (VS) Bregma: between 1.5 mm and 1.0 mm, Lateral: 1.8 mm, Depth: 3.8 mm, angle 10º
818 - (DMS) Bregma: between 1.5 mm and 0 mm, Lateral: 1.3 mm, Depth: 2.3 mm
819 - (DLS) Bregma: between 1.3 mm and -0.8 mm, Lateral: 3.0 mm, Depth: 2.3 mm
820
821 **Behavioral tasks**
822 All behavioral experiments were performed in custom-built behavioral rigs and controlled by a
823 NIDAQ board (National Instruments, TX) and Labview (National Instruments, TX), similar to a
824 previous study (Uchida and Mainen, 2003). Mice were trained to perform an odor-discrimination
825 task for water reward, similar to a study in rats (Uchida and Mainen, 2003) with several
826 modification. Mice initiated trials in a self-paced manner by poking a center port, which then
827 delivered an odor. Different odors were used in a pseudorandomized order from 3 different pure
828 chemicals (odor A, B and C) and mixtures of odor A and B with various ratios. Mice were
829 required to choose a left or right water port depending on dominant odor identity, odor A or B.
830 Correct choice was always rewarded by a drop of water. Odor C was never associated with
831 outcomes. To isolate cue- and water-related signals from potential motion artifacts in recording
832 and motor-related activity, mice were required to stay in an odor port for at least 1 s, and then to
833 stay in a water port for 1 s to get water reward. The inter-trial-interval was fixed at 7 s after water
834 onset in correct trials and at 9 s after any types of an error including violation of the stay
835 requirement, no choice within 5 s after odor port out, and multiple pokes of an odor port after
836 odor delivery. 1-Butanol, eugenol and cymene were diluted in 1/10 with mineral oil and
837 randomly assigned to odor A, B or C across animals. The odor-port assignment (left or right) was
838 held constant in a single animal.
839
840 Mice were first trained only with pure odors and with the same amounts of water reward (~6 ul).
841 After mice achieved greater than 90% accuracy, mice received a surgery for viral injection and
842 fiber implantation. Following a 1-week recovery period, mice received re-training and then,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mixtures of odor A and B (100/0, 90/10, 65/35, 35/65, 10/90, 0/100) were gradually introduced. After the accuracy of all the mixture odors achieved more than 50%, neuronal recording with fiber fluorometry was performed for 5 sessions. Subsequently, a task with different amounts of water was introduced. Mixtures of odor A and B (100/0, 65/35, 35/65, 0/100) but no odor C were used in this task. Each recording session started with 88-120 trials with an equal amount of water (~6 ul, the standard amount) in the first block to calibrate any potential bias on the day. In the second block, different amounts of reward were delivered in each water port. In order to make the water amounts unpredictable, one water port delivered big or medium size of water (2.2 and 0.8 times of the standard, ~13.2 and 4.8 µl, BIG side) in a pseudo-random order, and another water port delivered medium or small size of water (0.8 and 0.2 times of the standard, ~4.8 and 1.2 µl, SMALL side) in a pseudo-random order. Block 2 continued for 200 trials or until the end of recording sessions, whichever came earlier. A mouse performed 134.3 ± 3.4 (mean ± SEM) trials in block 2. The water condition (BIG or SMALL) was assigned to a left or right water port in a pseudo-random order across sessions. Recording was conducted for 40 min every other day to avoid potential bleaching. On days with no recording, animals were trained with pure odors A and B with the standard amount of water.

Fiber photometry

Fiber fluorometry (photometry) was performed as previously reported (Menegas et al., 2018) with a few modification. The optic fiber (400 µm diameter, Doric Lenses) allows chronic, stable, minimally disruptive access to deep brain regions and interfaces with a flexible patch cord (Doric Lenses, Canada) on the skull surface to simultaneously deliver excitation light (473 nm, Laserglow Technologies, Canada; 561 nm, Opto Engine LLC, UT) and collect GCaMP and tdTomato fluorescence emissions. Activity-dependent fluorescence emitted by cells in the vicinity of the implanted fiber's tip was spectrally separated from the excitation light using a dichroic, passed through a single band filter, and focused onto a photodetector connected to a current preamplifier (SR570, Stanford Research Systems, CA). During recording, optic fibers were connected to a magnetic patch cable (Doric Lesnses, MFP_400/430) which delivered excitation light (473 nm and 561 nm) and collected all emitted light. The emitted light was subsequently filtered using a 493/574 nm beam-splitter (Semrock, NY) followed by a 500 ± 20 nm (Chroma, VT) and 661 ± 20 nm (Semrock, NY) bandpass filters and collected by a

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

874 photodetector (FDS10x10 silicone photodiode, Thorlabs, NJ) connected to a current preamplifier
875 (SR570, Stanford Research Systems, CA). This preamplifier output a voltage signal which was
876 collected by a NIDAQ board (National Instruments, TX) and Labview software (National
877 Instruments, TX).
878
879 **Histology**
880 Mice were perfused using 4% paraformaldehyde and then brains were sliced into 100 µm thick
881 coronal sections using a vibratome and stored in PBS. Slices were then mounted in anti-fade
882 solution (VECTASHIELD anti-fade mounting medium, H-1000, Vector Laboratories, CA) and
883 imaged using a Zeiss Axio Scan Z1 slide scanner fluorescence microscope (Zeiss, Germany).
884
885 **Behavior analysis**
886 We fitted % of odor mixture (X) to % of choice left or choice BIG (µ) using generalized linear
887 model with logit link function in each animal as previously reported (Uchida and Mainen, 2003).
888 $\log(\mu/(1-\mu)) = Xb_1 + b_0$
889 We first fitted a control block (block 1) and a reward-manipulation block (block 2) separately to
890 examine difference of a slope, b_1 and a bias, $50\text{-}b_0/b_1$ of the curve. Next, to quantify shift of
891 choice bias, we fitted choice of block 1 and block 2 together with a fixed slope, by fitting odor
892 (X_1) and a block type (X_2=0 for block 1, X_2=1 for block 2) to choice.
893 $\log(\mu/(1-\mu)) = X_1b_1 + X_2b_2 + b_0$
894 Choice bias in block 2 was quantified choice bias as a lateral shift of the psychometric curve
895 equivalent to % mixture of odors, $50 - (b_0 + b_2)/ b_1$, which is a lateral shift compared to no bias,
896 and $b_0/b_1 - (b_0 + b_2)/ b_1$, which is a lateral shift compared to choice in block 1.
897
898 **GCaMP detection and analysis**
899 To synchronize behavioral events and fluorometry signals, TTL signals were sent every 10 s
900 from a computer that was used to control and record task events using Labview, to a NIDAQ
901 board that collects fluorometry voltage signals. GCaMP and tdTom signals were collected as
902 voltage measurements from current preamplifiers. Green and red signals were cleaned by
903 removing 60Hz noise with bandstop FIR filter 58-62Hz and smoothing with moving average of
904 signals in 50ms. The global change within a session was normalized using a moving median of

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

100s. Then, the correlation between green and red signals during ITI was examined by linear regression. If the correlation is significant ($p<0.05$), fitted tdTom signals were subtracted from green signals.

Responses were calculated by subtracting the average baseline activity from the average activity of the target window. Unless specified otherwise, odor responses were calculated by averaging activity from 1-0 s before odor port out (before choice) minus the average activity from the baseline period (1-0.2 s before odor onset). Responses after choice were calculated by averaging activity from 0-1 s after water port in minus the same baseline. Outcome responses were calculated by averaging activity from 0-1 s after water onset minus the same baseline. When comparing activity before and after water onset, average activity in 1-0.2 s before water onset was used as baseline. To normalize GCaMP signals across sessions within an animal, GCaMP signals were divided by average of peak responses during 1 s after odor onset in all the successful trials in the session. Z-scores of the signals were obtained using mean and standard deviation of signals in all the choice trials (from 2 s before odor onset to 6 s after odor onset) in each animal.

We built a regularized linear regression to fit cosine kernels (Park et al., 2014) (width of 200 ms, interval of 40 ms) to the activity of dopamine axons in each animal. We used down-sampled (every 20 ms) responses in all valid choice trials (trials with >1s odor sampling time and any choice, -1 to 7 s from odor onset) for the model fitting. We used 4 different time points to lock kernels: odor onset ("odor"), odor port out ("movement"), water port in ("choice"), and water onset ("water"). Odor kernels consist of 4 types of kernels: "base" kernels to span -960 to 200 ms from odor onset in all trials, and "pure big" kernels in trials with a pure odor associated with big/medium water, "pure small" kernels in trials with a pure odor associated with medium/small water, and "mixture" kernels in trials with a mixture odor to span 0-1600 ms from odor onset. Movement kernels consist of 2 types of kernels: "contra turn" kernels in trials with choice contra-lateral to the recording site, and "ipsi turn" kernels in trials with choice ipsi-lateral to the recording site to span -1000 to 1200 ms from when a mouse exited an odor port. Choice kernels consist of 3 types of kernels: "correct big" kernels in trials with correct choice of medium/small water and "correct small" kernels in trials with correct choice of medium/small water to span -

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

400 to 1200 ms from when a mouse entered a water port (water port in), and "error" kernels in trials with choice error to span -400 to 5200 ms from water port in. Water kernels consist of 4 types of kernels: "big water" kernels for big size of water, "medium water big side" kernels for medium size of water at a water port of big/medium water, "medium water small side" kernels for medium size of water at a water port of medium/small water, and "small water" for small size of water to span 0-4200 ms after water onset. All the kernels were fitted to responses using linear regression with Lasso regularization with 10-fold cross validation. Regularization coefficient lambda was chosen so that cross-validation error is minimum plus one standard deviation. % explained by a model was expressed as reduction of a variance in the residual responses compared to the original responses. Contribution of each component in the model was measured by reduction of a deviance compared to a reduced model excluding the component.

We estimated response function to water in dopamine axons with linear regression with power function in each animal.

$$r = k(R^{\alpha} + c1 * S + c2)$$

where r is the dopamine axon response to water, R is the water amount, S is SMALL side (S=1 when water was delivered at SMALL side, S=0 otherwise). There are 4 different conditions, responses to big and medium water at a port of BIG side, and to medium and small water at a port of SMALL side. We first optimized α by minimizing average of residual sum of squares for each animal and then applied α =0.7 for all the animals to obtain other parameters, k, c1, and c2. The response function was drawn with R as x-axis and r as y-axis. The amount of water to which dopamine axons do not respond under expectation of BIG or SMALL water was estimated by getting a crossing point of the obtained response function where the value is 0 (a zero-crossing point). The distribution of zero-crossing points was examined by linear regression of zero-crossing values against anatomical locations (anterior-posterior, dorsal-ventral, and medial-lateral). To visualize zero crossing points on the atlas, zero-crossing values were fitted against anatomical locations with interaction terms using linear regression with elastic net regularization (α=0.1) with 3-fold cross validation. The constructed map was sliced at a coronal plane Bregma +0.7 and overlaid on an atlas (Paxinos and Franklin, 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To visualize activity pattern in multiple time windows at the same time, we stretched activity in each trial to standard windows. Standard windows from odor onset to odor poke out, and from odor poke out to water poke in, were determined by median reaction time and median movement time for each animal. For average plots of multiple animals, windows were determined by the average of median reaction times and of median movement times in all animals. The number of 100ms bins in each time window was determined by dividing median reaction time and median movement time by 100. Dopamine responses in the window were divided into the bin number and the average response in each bin was stretched to 100ms. The stretched activity patterns were used only for visualization, and all the statistical analyses were performed using original responses.

Estimation of state values and TD errors using simulations

To examine how the value and RPE may change within a trial, we employed a Monte-Carlo approach to simulate animal's choices at a steady state (i.e. after the animal learned the task). We used a Monte-Caro approach to obtain the *ground truth* state values as the animal progresses through task events without assuming a specific learning algorithm, under the assumption that the animal has learned the task. After obtaining the state values, we computed TD errors over the obtained state values.

Model architecture

We considered two types of models. The variability and errors in choice in psychophysical performance can arise from at least two noise sources; noise in the variability in the process of estimating sensory inputs (perceptual noise) and noise in the process of selecting an action (decision noise). The first model contained only perceptual noise (Green and Swets, 1966), and the second model contained both perceptual and decision noise.

These models had different "states" considering N_S subjective odors (N_S = 60 or 4 discrete states), choice (BIG versus SMALL), and different timing (inter-trial interval, odor port entry, odor presentation, choice, water port in, waiting for reward, and receiving feedback/outcome) (circles in Figure 7A).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We assumed N_S possible subjective odor states (O') which comprise SubOdor1 and SubOdor2 states. We assumed that, in each trial, an internal estimate of the stimulus or a "subjective odor" (O') was obtained by adding a noise to the presented odor stimulus (O) (one of the 4 mixtures of Odor A and B; 100/0, 65/35, 35/65, 0/100) (Figure 7A-C). In the model, the probability of falling on a given subjective odor state (O') is calculated using a Gaussian distribution centering on the presented odor (O) with the standard deviation, σ. We considered two successive states for subjective odor states in order to reflect a relatively long duration before an odor port exit.

As in the behavioral paradigm, whether the model receives a reward or not was determined solely by whether the presented odor (O) instructed the BIG side or SMALL side. Each subjective odor state contains cases when the presented odor (O) is consistent or congruent with the subjective odor (O'). For each subjective odor state, the probability of receiving a reward after choosing the BIG side, $p(BIG\ is\ correct) = f_B$, can be calculated as the fraction of cases when the presented odors instructed the BIG side. Conversely, the probability of reward after choosing the SMALL side is $p(SMALL\ is\ correct) = f_S = 1 - f_B$. Note that neither f_B nor f_S depends on reward size manipulations (as will be discussed later, the animal's choices will be dependent on reward size manipulations).

Action selection

For each subjective odor, the model chose either the BIG or the SMALL side based on the value of choosing the BIG or SMALL side (V_B and V_S respectively, equivalent to the state value of the next state after committing to choose the BIG or SMALL side; see below for how V_B and V_S were obtained). In the first model which contains only perceptual noise, the side that is associated with a larger value is chosen. In the second model which contains both perceptual and decision noise, a choice is made by transforming V_B and V_S into the probability of choosing a given option using a sigmoidal function (e.g. Boltzmann distribution) (Sutton and Barto, 2011). In the softmax, the probabilities of choosing the BIG and SMALL side (P_B, P_S) are given, respectively, by,

$$P_B = \frac{e^{(V_B/(V_B+V_S))/\tau}}{e^{(V_B/(V_B+V_S))/\tau} + e^{(V_S/(V_B+V_S))/\tau}}$$

$$P_S = 1 - P_B$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We also tested other stochastic decision rules such as Herrnstein's matching law (Herrnstein, 1961) or ε-greedy exploration (randomly selecting an action in a certain fraction [ε] of trials) (Sutton and Barto, 2011). In Herrnstein's matching law, the probability of choosing the BIG side is given by,

$$P_B = \frac{V_S}{V_S + V_B}$$

The perceptual noise and a set of decision rule determine the behavioral performance of the model. The first model has only one free parameter, σ. The second model has one or no additional parameter (τ for softmax, or ε, for ε-greedy; no additional parameter for matching). We first obtained the best fit parameter(s) based on the behavioral performance of all animals (the average performance in Block 2; i.e. Figure 1C, orange) by minimizing the mean squared errors in the psychometric curves.

For the first model, the best fit σ was 21% Odor. We also tested with σ of 5%, and the TD error dynamic was qualitatively similar. For the second model using the softmax rule, the best fit τ was 0.22 while σ was 18% Odor.

State values

The state value for each state was obtained as the weighted sum of expected values of available options which was computed by multiplying expected values of the option with probability of an option in the next step.

Outcome2 state represents the timing when the animal recognizes the amount of water. The state value is given by the amount of water that the model received (big, medium, small),

$$V_b = 2.2^{\alpha}$$

$$V_m = 0.8^{\alpha}$$

$$V_s = 0.2^{\alpha}$$

where the exponent $\alpha = 0.7$ makes the value function a concave function of reward amounts, similar to the fitting analysis of the fluorometry data (Figure 4C). Using $\alpha = 1$ (i.e. a linear function) did not change the results.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ourcome1 state, or Water/No-water states (W and N, respectively) represent when the animal noticed the presence or absence of reward, respectively, but not the amount of reward. The value of a W (Water) state was defined by the average value of the next states. At the BIG side,

$$V_{WB} = (V_b + V_m)/2$$

whereas at the SMALL side,

$$V_{WS} = (V_m + V_s)/2$$

The values of N (No-water) states at the BIG and SMALL side are zero,

$$V_{NB} = 0$$

$$V_{NS} = 0$$

WaterPort1 and WaterPort2 states represent when the animal entered and stayed in the water port, respectively. The state value was obtained separately for the BIG and SMALL side. The value of choosing the BIG and SMALL sides is given by weighted sum of the values of the next states (V_{WB}, V_{NB}, V_{WS}, V_{NS}). The probabilities of transiting to the W and N states are given by the probability of receiving a reward given the choice (BIG or SMALL). As discussed above, these probabilities are given by f_B and f_S, respectively. Thus,

$$V_B = f_B \cdot V_{WB}$$

$$V_S = f_S \cdot V_{WS}$$

We considered two successive states for WaterPort states to reflect a relatively long duration before receiving feedback/outcome. The two successive states had the same state values.

SubOdor1 and SubOdor2 states represent when the animal obtained a subjective odor (O') and before making a choice. The model chooses the BIG or SMALL side with the probability of P_B and P_S, respectively, as defined above. Therefore, the state value of WaterPort1 and WaterPort2 was defined by the weighted sum of the values of the next states (V_B and V_S),

$$V_{O'} = P_B V_B + P_S V_S$$

The two successive states had the same state values.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

OdorOn state represents when the animal recognized the presentation of an odor but before recognizing the identity of that odor. The state value of the OdorOn state is defined by the weighted sum of the values of the next states (SubOdor1).

ITI state represents when the animal is in the inter-trial interval (i.e. before odor presentation). The value of ITI state was set to zero.

TD errors

After obtaining state values at each state, we then computed TD errors using a standard definition of TD error which is the difference between the state values at consecutive time points plus received rewards at each time step (Sutton and Barto, 1987). For simplicity, a discounting factor was set to 1 (no discounting).

Invalid trials

We also tested the effect of including invalid trials. At water acquisition, we included failures (20% of trials, value 0) where a mouse did not fulfil the requirement of odor poke duration (short odor poke), but did indicate a choice. At an odor port, failures resulted from multiple pokes of odor port (4% of trials), and a short odor poke (14% of trials). Values for these failures were set to 0. Existence or omission of these failures in models did not change the conclusion.

Randomization, blinding, and data exclusion

Chemicals were randomly assigned to an odor cue. Trial types (odors) were pseudorandomized in a block. Session types were pseudorandomized in a recording schedule. Animals were randomly assigned to a recording location. The experimenter did not know location of recording until the recording schedule was completed. No animals were excluded from the study: all analysis includes data from all animals. No trials were excluded from statistical analyses. To visualize average activity pattern in a stretched time-window, outlier trials (maximum, minimum or average activity of a trial is outside of 3 × standard deviation of maximum, minimum or average activity of all the trials) were excluded.

Statistical analyses

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data analysis was performed using custom software written in Matlab (MathWorks, Natick, MA, USA). All code used for analysis is available on request. All statistical tests were two-sided. For statistical comparisons of the mean, we used one-way ANOVA and two-sample Student's t tests, unless otherwise noted. Paired t tests were conducted when the same mouse's neural activity was being compared across different conditions or different time windows. The significance level was corrected for multiple comparisons using Holm–Sidak's tests unless otherwise indicated. All error bars in the figures are s.e.m. In boxplots, the edges of the boxes are the 25th and 75th percentiles (q1 and q3, respectively), and the whiskers extend to the most extreme data points not considered outliers. Points are drawn as outliers if they are larger than q3+1.5×(q3-q1) or q1-1.5×(q3-q1). Individual data points were overlaid on boxplots to compare striatal areas.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author Contributions:

ITK and MWU designed experiments and collected and analyzed data. HM performed pilot experiments. The results were discussed and interpreted by ITK, NU and MWU. ITK, NU and MWU wrote the paper.

Acknowledgments

We thank Ju Tian, William Menegas, HyungGoo Kim and Takahiro Yamaguchi for technical assistance, Kristen Fang, Grace Chang, Melissa Yamada and Sakura Ikeda for assistance in animal training and histology, and all lab members for discussion. We also thank V. Jayaraman, R. Kerr, D. Kim, L. Looper, and K. Svoboda from the GENIE Project, Janelia Farm Research Campus, Howard Hughes Medical Institute for AAV-FLEX-GCaMP7f. This work was supported by National Institute of Mental Health R01MH095953, R01MH101207, R01MH110404, R01NS108740 (NU); and Japan Society for the Promotion of Science, Japan Science and Technology Agency (HM, ITK).

Declaration of Interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCE

Adler, W.T., and Ma, W.J. (2018). Limitations of Proposed Signatures of Bayesian Confidence. Neural Comput. *30*, 3327–3354.

Bäckman, C.M., Malik, N., Zhang, Y., Shan, L., Grinberg, A., Hoffer, B.J., Westphal, H., and Tomac, A.C. (2006). Characterization of a mouse strain expressing Cre recombinase from the 3' untranslated region of the dopamine transporter locus. Genes. N. Y. N 2000 *44*, 383–390.

Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407–419.

Balleine, B.W., and O'Doherty, J.P. (2010). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. Neuropsychopharmacology *35*, 48–69.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141.

Brown, H.D., McCutcheon, J.E., Cone, J.J., Ragozzino, M.E., and Roitman, M.F. (2011). Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. Eur. J. Neurosci. *34*, 1997–2006.

Chang, C.Y., Gardner, M.P.H., Conroy, J.C., Whitaker, L.R., and Schoenbaum, G. (2018). Brief, But Not Prolonged, Pauses in the Firing of Midbrain Dopamine Neurons Are Sufficient to Produce a Conditioned Inhibitor. J. Neurosci. *38*, 8822–8830.

Chen, T.-W., Li, N., Daie, K., and Svoboda, K. (2017). A map of anticipatory activity in mouse motor cortex. Neuron *94*, 866–879.

Clark, J.J., Hollon, N.G., and Phillips, P.E. (2012). Pavlovian valuation systems in learning and decision making. Curr. Opin. Neurobiol. *22*, 1054–1061.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85–88.

Cox, J., and Witten, I.B. (2019). Striatal circuits for reward learning and decision-making. Nat. Rev. Neurosci. *20*, 482–494.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. Nature 1–5.

Dana, H., Sun, Y., Mohar, B., Hulse, B.K., Kerlin, A.M., Hasseman, J.P., Tsegaye, G., Tsang, A., Wong, A., Patel, R., et al. (2019). High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. Nat. Methods *16*, 649–657.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1186 Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal
1187 and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

1188 Dayan, P., and Berridge, K.C. (2014). Model-based and model-free Pavlovian reward learning:
1189 Revaluation, revision, and revelation. Cogn. Affect. Behav. Neurosci. *14*, 473–492.

1190 Dezfouli, A., and Balleine, B.W. (2012). Habits, action sequences and reinforcement learning.
1191 Eur. J. Neurosci. *35*, 1036–1051.

1192 Dickinson, A., and Weiskrantz, L. (1985). Actions and habits: the development of behavioural
1193 autonomy. Philos. Trans. R. Soc. Lond. B Biol. Sci. *308*, 67–78.

1194 Dolan, R.J., and Dayan, P. (2013). Goals and Habits in the Brain. Neuron *80*, 312–325.

1195 Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H.J., Ornelas, S., Koay, S.A., Thiberge,
1196 S.Y., Daw, N.D., Tank, D.W., et al. (2019). Specialized coding of sensory, motor and cognitive
1197 variables in VTA dopamine neurons. Nature *570*, 509–513.

1198 Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common
1199 response function for reward prediction error. Nat. Neurosci. *19*, 479–486.

1200 Evans, R.C., Zhu, M., and Khaliq, Z.M. (2017). Dopamine inhibition differentially controls
1201 excitability of substantia nigra dopamine neuron subpopulations through T-type calcium
1202 channels. J. Neurosci. *37*, 3704–3720.

1203 Farassat, N., Costa, K.M., Stojanovic, S., Albert, S., Kovacheva, L., Shin, J., Egger, R., Somayaji,
1204 M., Duvarci, S., and Schneider, G. (2019). In vivo functional diversity of midbrain dopamine
1205 neurons within identified axonal projections. Elife *8*.

1206 Gerfen, C.R., and Surmeier, D.J. (2011). Modulation of Striatal Projection Systems by Dopamine.
1207 Annu. Rev. Neurosci. *34*, 441–466.

1208 Graybiel, A.M. (2008). Habits, Rituals, and the Evaluative Brain. Annu. Rev. Neurosci. *31*, 359–
1209 387.

1210 Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics (Wiley New
1211 York).

1212 Hangya, B., Sanders, J.I., and Kepecs, A. (2016). A Mathematical Framework for Statistical
1213 Decision Confidence. Neural Comput. *28*, 1840–1858.

1214 Hart, A.S., Rutledge, R.B., Glimcher, P.W., and Phillips, P.E. (2014). Phasic dopamine release in
1215 the rat nucleus accumbens symmetrically encodes a reward prediction error term. J. Neurosci.
1216 *34*, 698–704.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Herrnstein, R.J. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement.

Hikosaka, O., Rand, M.K., Miyachi, S., and Miyashita, K. (1995). Learning of sequential movements in the monkey: process of learning and retention of memory. J. Neurophysiol. *74*, 1652–1661.

Hirokawa, J., Vaughan, A., Masset, P., Ott, T., and Kepecs, A. (2019). Frontal cortex neuron types categorically encode single decision variables. Nature *576*, 446–451.

Hong, S., Jhou, T.C., Smith, M., Saleem, K.S., and Hikosaka, O. (2011). Negative reward signals from the lateral habenula to dopamine neurons are mediated by rostromedial tegmental nucleus in primates. J. Neurosci. *31*, 11457–11471.

Howe, M.W., and Dombeck, D.A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. Nature *535*, 505–510.

Hunnicutt, B.J., Jongbloets, B.C., Birdsong, W.T., Gertz, K.J., Zhong, H., and Mao, T. (2016). A comprehensive excitatory input map of the striatum reveals novel functional organization. ELife *5*.

Ilango, A., Kesner, A.J., Keller, K.L., Stuber, G.D., Bonci, A., and Ikemoto, S. (2014). Similar Roles of Substantia Nigra and Ventral Tegmental Dopamine Neurons in Reward and Aversion. J. Neurosci. *34*, 817–822.

Jhou, T.C., Geisler, S., Marinelli, M., Degarmo, B.A., and Zahm, D.S. (2009a). The mesopontine rostromedial tegmental nucleus: a structure targeted by the lateral habenula that projects to the ventral tegmental area of Tsai and substantia nigra compacta. J. Comp. Neurol. *513*, 566–596.

Jhou, T.C., Fields, H.L., Baxter, M.G., Saper, C.B., and Holland, P.C. (2009b). The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. Neuron *61*, 786–800.

de Jong, J.W., Afjei, S.A., Pollak Dorocic, I., Peck, J.R., Liu, C., Kim, C.K., Tian, L., Deisseroth, K., and Lammel, S. (2019). A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. Neuron *101*, 133-151.e7.

Kamin, L.J. (1969). Predictability, surprise, attention and conditioning. Punishm. Aversive Behav.

Keiflin, R., Pribut, H.J., Shah, N.B., and Janak, P.H. (2019). Ventral Tegmental Dopamine Neurons Participate in Reward Identity Predictions. Curr. Biol. *29*, 93-103.e3.

Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature *455*, 227–231.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1250 Kim, H.F., Ghazizadeh, A., and Hikosaka, O. (2015). Dopamine Neurons Encoding Long-Term
1251 Memory of Object Value for Habitual Behavior. Cell *163*, 1165–1175.

1252 Kudo, Y., Akita, K., Nakamura, T., Ogura, A., Makino, T., Tamagawa, A., Ozaki, K., and Miyakawa,
1253 A. (1992). A single optical fiber fluorometric device for measurement of intracellular Ca2+
1254 concentration: its application to hippocampal neurons in vitro and in vivo. Neuroscience *50*,
1255 619–625.

1256 Lak, A., Nomoto, K., Keramati, M., Sakagami, M., and Kepecs, A. (2017). Midbrain dopamine
1257 neurons signal belief in choice accuracy during a perceptual decision. Curr. Biol. *27*, 821–832.

1258 Lak, A., Hueske, E., Hirokawa, J., Masset, P., Ott, T., Urai, A.E., Donner, T.H., Carandini, M.,
1259 Tonegawa, S., Uchida, N., et al. (2020a). Reinforcement biases subsequent perceptual decisions
1260 when confidence is low, a widespread behavioral phenomenon. ELife *9*, e49834.

1261 Lak, A., Okun, M., Moss, M.M., Gurnani, H., Farrell, K., Wells, M.J., Reddy, C.B., Kepecs, A.,
1262 Harris, K.D., and Carandini, M. (2020b). Dopaminergic and prefrontal basis of learning from
1263 sensory confidence and reward value. Neuron *105*, 700–711.

1264 Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B., and Roeper, J. (2008). Unique Properties of
1265 Mesoprefrontal Neurons within a Dual Mesocorticolimbic Dopamine System. Neuron *57*, 760–
1266 773.

1267 Lee, K., Claar, L.D., Hachisuka, A., Bakhurin, K.I., Nguyen, J., Trott, J.M., Gill, J.L., and
1268 Masmanidis, S.C. (2020). Temporally restricted dopaminergic control of reward-conditioned
1269 movements. Nat. Neurosci. *23*, 209–216.

1270 Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K.,
1271 Malenka, R.C., Luo, L., Tomer, R., et al. (2015). Intact-Brain Analyses Reveal Distinct Information
1272 Carried by SNc Dopamine Subcircuits. Cell *162*, 635–647.

1273 Li, H., Vento, P.J., Parrilla-Carrero, J., Pullmann, D., Chao, Y.S., Eid, M., and Jhou, T.C. (2019).
1274 Three Rostromedial Tegmental Afferents Drive Triply Dissociable Aspects of Punishment
1275 Learning and Aversive Valence Encoding. Neuron *104*, 987-999.e4.

1276 Lloyd, K., and Dayan, P. (2016). Safety out of control: dopamine and defence. Behav. Brain
1277 Funct. BBF *12*, 15.

1278 Madisen, L., Zwingman, T.A., Sunkin, S.M., Oh, S.W., Zariwala, H.A., Gu, H., Ng, L.L., Palmiter,
1279 R.D., Hawrylycz, M.J., Jones, A.R., et al. (2010). A robust and high-throughput Cre reporting and
1280 characterization system for the whole mouse brain. Nat. Neurosci. *13*, 133–140.

1281 Malvaez, M., and Wassum, K.M. (2018). Regulation of habit formation in the dorsal striatum.
1282 Curr. Opin. Behav. Sci. *20*, 67–74.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1283 Matsumoto, M., and Hikosaka, O. (2007). Lateral habenula as a source of negative reward
1284 signals in dopamine neurons. Nature *447*, 1111–1115.

1285 Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey
1286 positive and negative motivational signals. Nature *459*, 837–841.

1287 Matsuzaka, Y., Picard, N., and Strick, P.L. (2007). Skill Representation in the Primary Motor
1288 Cortex After Long-Term Practice. J. Neurophysiol. *97*, 1819–1832.

1289 Menegas, W., Bergan, J.F., Ogawa, S.K., Isogai, Y., Umadevi Venkataraju, K., Osten, P., Uchida,
1290 N., and Watabe-Uchida, M. (2015). Dopamine neurons projecting to the posterior striatum
1291 form an anatomically distinct subclass. ELife *4*, e10032.

1292 Menegas, W., Babayan, B.M., Uchida, N., and Watabe-Uchida, M. (2017). Opposite initialization
1293 to novel cues in dopamine signaling in ventral and posterior striatum in mice. ELife *6*.

1294 Menegas, W., Akiti, K., Amo, R., Uchida, N., and Watabe-Uchida, M. (2018). Dopamine neurons
1295 projecting to the posterior striatum reinforce avoidance of threatening stimuli. Nat. Neurosci.
1296 *21*, 1421–1430.

1297 Miller, K.J., Shenhav, A., and Ludvig, E.A. (2019). Habits without values. Psychol. Rev. *126*, 292–
1298 311.

1299 Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic
1300 dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

1301 Morris, A., and Cushman, F. (2019). Model-Free RL or Action Sequences? Front. Psychol. *10*.

1302 O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R.J. (2004).
1303 Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. Science *304*,
1304 452–454.

1305 Ölveczky, B.P. (2011). Motoring ahead with rodents. Curr. Opin. Neurobiol. *21*, 571–578.

1306 Oyama, K., Hernádi, I., Iijima, T., and Tsutsui, K.-I. (2010). Reward Prediction Error Coding in
1307 Dorsal Striatal Neurons. J. Neurosci. *30*, 11447–11457.

1308 Park, I.M., Meister, M.L.R., Huk, A.C., and Pillow, J.W. (2014). Encoding and decoding in parietal
1309 cortex during sensorimotor decision-making. Nat. Neurosci. *17*, 1395–1403.

1310 Parker, N.F., Cameron, C.M., Taliaferro, J.P., Lee, J., Choi, J.Y., Davidson, T.J., Daw, N.D., and
1311 Witten, I.B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons
1312 depends on striatal target. Nat. Neurosci. *19*, 845–854.

1313 Paxinos, G., and Franklin, K.B.J. (2019). Paxinos and Franklin's the Mouse Brain in Stereotaxic
1314 Coordinates (Academic Press).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev. *87*, 532.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

Rausch, M., and Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. PLOS Comput. Biol. *15*, e1007456.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Class. Cond. II Curr. Res. Theory *2*, 64–99.

Robbins, T.W., and Costa, R.M. (2017). Habits. Curr. Biol. *27*, R1200–R1206.

Rorie, A.E., Gao, J., McClelland, J.L., and Newsome, W.T. (2010). Integration of Sensory and Reward Information during Perceptual Decision-Making in Lateral Intraparietal Cortex (LIP) of the Macaque Monkey. PLOS ONE *5*, e9308.

Sajad, A., Godlove, D.C., and Schall, J.D. (2019). Cortical microcircuitry of performance monitoring. Nat. Neurosci. *22*, 265–274.

Samejima, K., and Doya, K. (2007). Multiple Representations of Belief States and Action Values in Corticobasal Ganglia Loops. Ann. N. Y. Acad. Sci. *1104*, 213–228.

Saunders, B.T., Richard, J.M., Margolis, E.B., and Janak, P.H. (2018). Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties. Nat. Neurosci. *21*, 1072–1083.

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

da Silva, J.A., Tecuapetla, F., Paixão, V., and Costa, R.M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. Nature *554*, 244–248.

Smith, K.S., and Graybiel, A.M. (2016). Habit formation. Dialogues Clin. Neurosci. *18*, 33–43.

Stuphorn, V., Taylor, T.L., and Schall, J.D. (2000). Performance monitoring by the supplementary eye field. Nature *408*, 857–860.

Suri, R.E., and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience *91*, 871–890.

Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. Mach. Learn. *3*, 9–44.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sutton, R.S., and Barto, A.G. (1987). A temporal-difference model of classical conditioning. In Proceedings of the Ninth Annual Conference of the Cognitive Science Society, (Seattle, WA), pp. 355–378.

Sutton, R.S., and Barto, A.G. (2011). Reinforcement learning: An introduction.

Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning, second edition: An Introduction (MIT Press).

Thorndike, E.L. (1932). The fundamentals of learning (New York, NY, US: Teachers College Bureau of Publications).

Tian, J., Huang, R., Cohen, J.Y., Osakada, F., Kobak, D., Machens, C.K., Callaway, E.M., Uchida, N., and Watabe-Uchida, M. (2016). Distributed and Mixed Information in Monosynaptic Inputs to Dopamine Neurons. Neuron *91*, 1374–1389.

Uchida, N., and Mainen, Z.F. (2003). Speed and accuracy of olfactory discrimination in the rat. Nat. Neurosci. *6*, 1224–1229.

Wang, A.Y., Miura, K., and Uchida, N. (2013). The dorsomedial striatum encodes net expected return, critical for energizing performance vigor. Nat. Neurosci. *16*, 639–647.

Watabe-Uchida, M., and Uchida, N. (2018). Multiple dopamine systems: Weal and woe of dopamine. In Cold Spring Harbor Symposia on Quantitative Biology, (Cold Spring Harbor Laboratory Press), pp. 83–95.

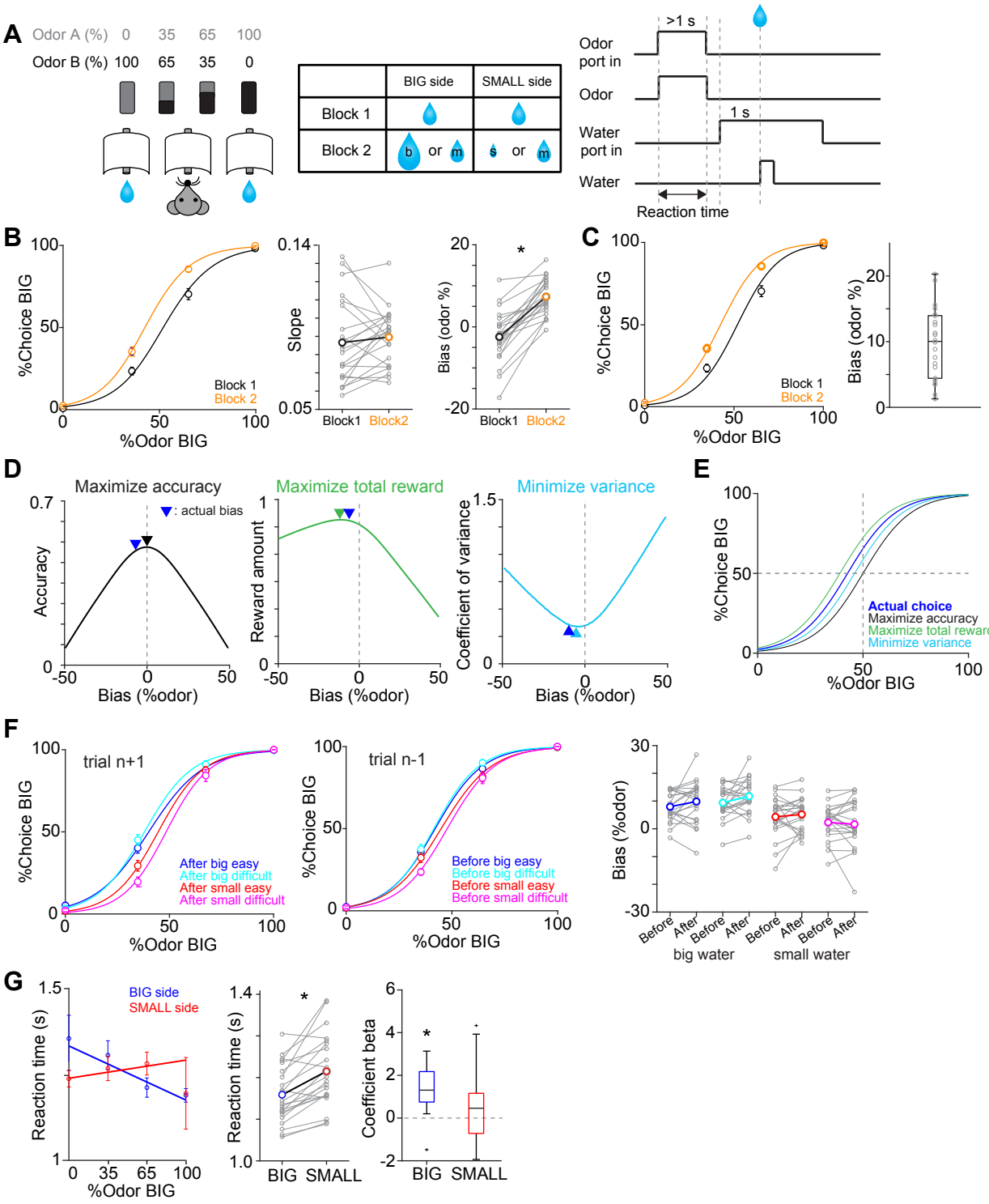
Yetnikoff, L., Lavezzi, H.N., Reichard, R.A., and Zahm, D.S. (2014). An update on the connections of the ventral mesencephalic dopaminergic complex. Neuroscience *282*, 23–48.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. *19*, 181–189.

Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005). The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1

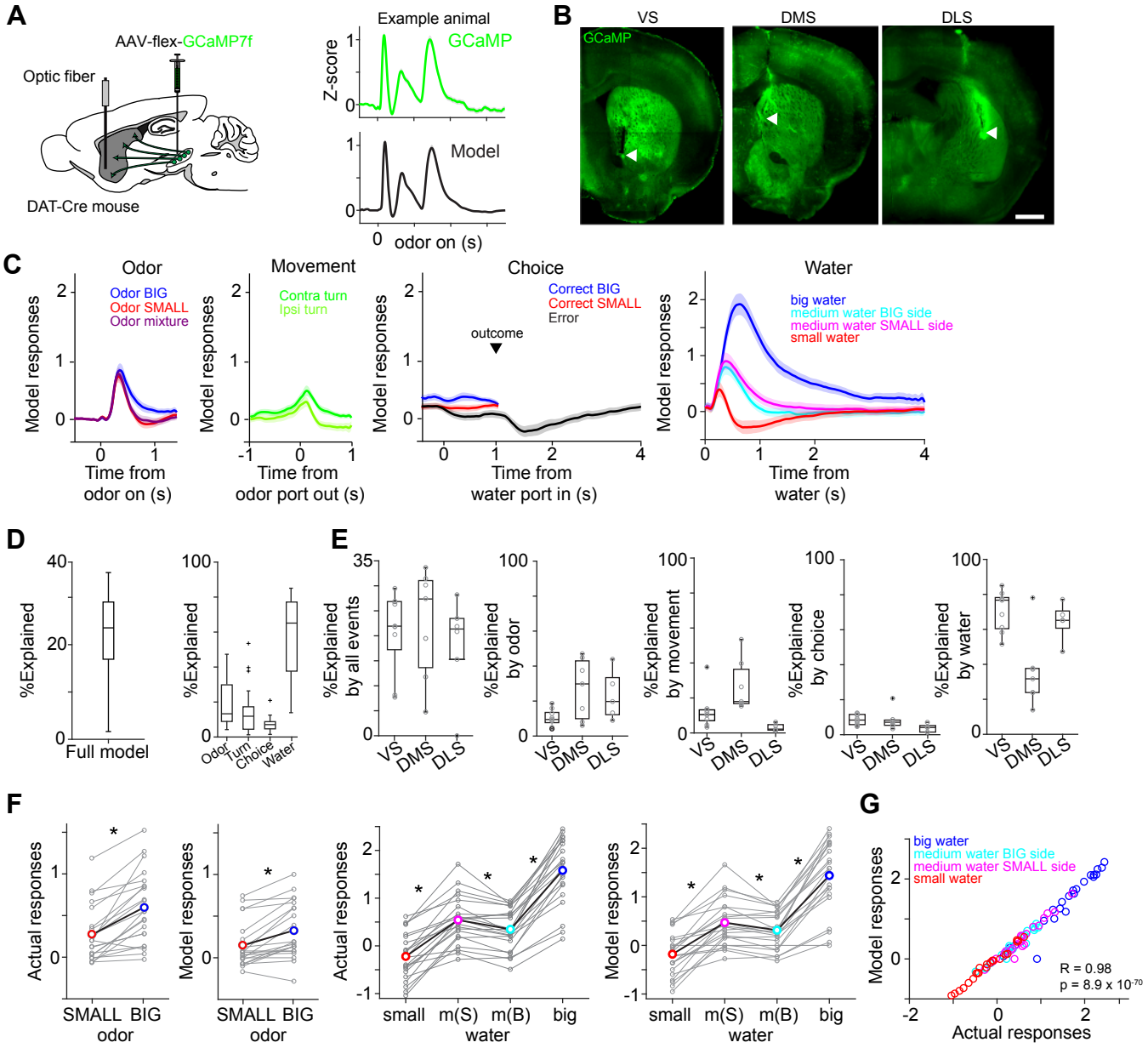


bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1. **Perceptual choice paradigm with probabilistic reward conditions** (A) A mouse discriminated a dominant odor in odor mixtures that indicates water availability in either the left or right water port. Correct choice was rewarded by a drop of water. In each session, an equal amount of water was assigned at both water ports in the first block, and in the second block, big/medium water (50% 50%, randomized) was assigned at one water port (BIG side) and medium/small water (50% 50%, randomized) was assigned at another port (SMALL side). The BIG or SMALL side was assigned to a left or right water port in a pseudorandom order across sessions. (B) Left, % of choice of the BIG side in block 1 and 2 (mean ± SEM) and the average psychometric curve for each block. Center, slope of the psychometric curve. Right, choice bias at 50/50 choice, expressed as 50 - odor (%). (C) Left, % of choice of the BIG side in block 1 and 2 (mean ± SEM) and the average psychometric curve with a fixed slope across blocks. Right, all the animals showed choice bias toward BIG side in block 2 compared to block 1. The choice bias was expressed by a lateral shift of a psychometric curve with a fixed slope across blocks. (D) Average reward amounts, accuracy, and coefficients of variance were examined with different levels of choice bias with a fixed slope (average slope of all animals). (E) Optimal choice patterns with different strategies in D (bias -11, 0, and -4, respectively) and the actual average choice pattern (mean bias -7.3). (F) Trial-by-trial choice updating was examined by comparing choice bias before (center, trial n-1) and after (left, trial n+1) specific trial types. Choice updating in one trial was not significant for reward acquisition of either small or big water in easy or difficult trials (right). (G) Left, animal's reaction time was modulated by odor types. Center, for easy trials (pure odors, correct choice), reaction time was shorter when animals chose the BIG side ($p=2.7\times10^{-5}$). Right, the reaction time was negatively correlated with sensory evidence for choice of the BIG side ($p=1.2\times10^{-4}$), whereas the modulation was not significant for choice of the SMALL side ($p=0.13$). n = 22 animals.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. Dopamine axons in the striatum show characteristics of RPE (A) AAV-flex-GCaMP7f was injected in VTA and SNc, and dopamine axon activity was measured with an optic fiber inserted in the striatum. Right top, dopamine axon activity in all the valid trials (an animal chose an either water port after wait for the required stay time) in an example animal, aligned at odor onset (mean ± SEM). Right bottom, a fitted model of the same animal (mean ± SEM). (B) Location of an optic fiber in example animals. Arrow heads, tips of fibers. Green, GCaMP7f. Bar = 1 mm (C) Odor-, movement-, choice-, and water-locked components in the model of all the animals (mean ± SEM). (D) Contribution of each component in the model was measured by reduction of deviance in the full model compared to a reduced model excluding the component. (E) Contribution of each component in the model in each animal group. (F) Left, comparison of dopamine axon responses to an odor cue that instructs to choose BIG and SMALL side in easy trials (pure odor, correct choice, -1-0 s before odor port out). $p=5.0\times10^{-6}$ for actual signals and $p=7.4\times10^{-5}$ for models. Right, comparison of dopamine axon responses to different sizes of water (big versus medium water with BIG expectation, and medium versus small water with SMALL expectation) and to medium water with different expectation (BIG versus SMALL expectation) (0.3-1.3 s after water onset). $p=1.2\times10^{-11}$, $p=3.8\times10^{-9}$ and $p=3.9\times10^{-4}$, respectively for actual signals, and $p=1.0\times10^{-9}$, $p=1.0\times10^{-7}$, and $p=0.0031$, respectively for models. n=22 animals. m(B), medium water with BIG expectation; m(S), medium water with SMALL expectation. (G) Comparison between actual dopamine axon responses and model responses to water.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3

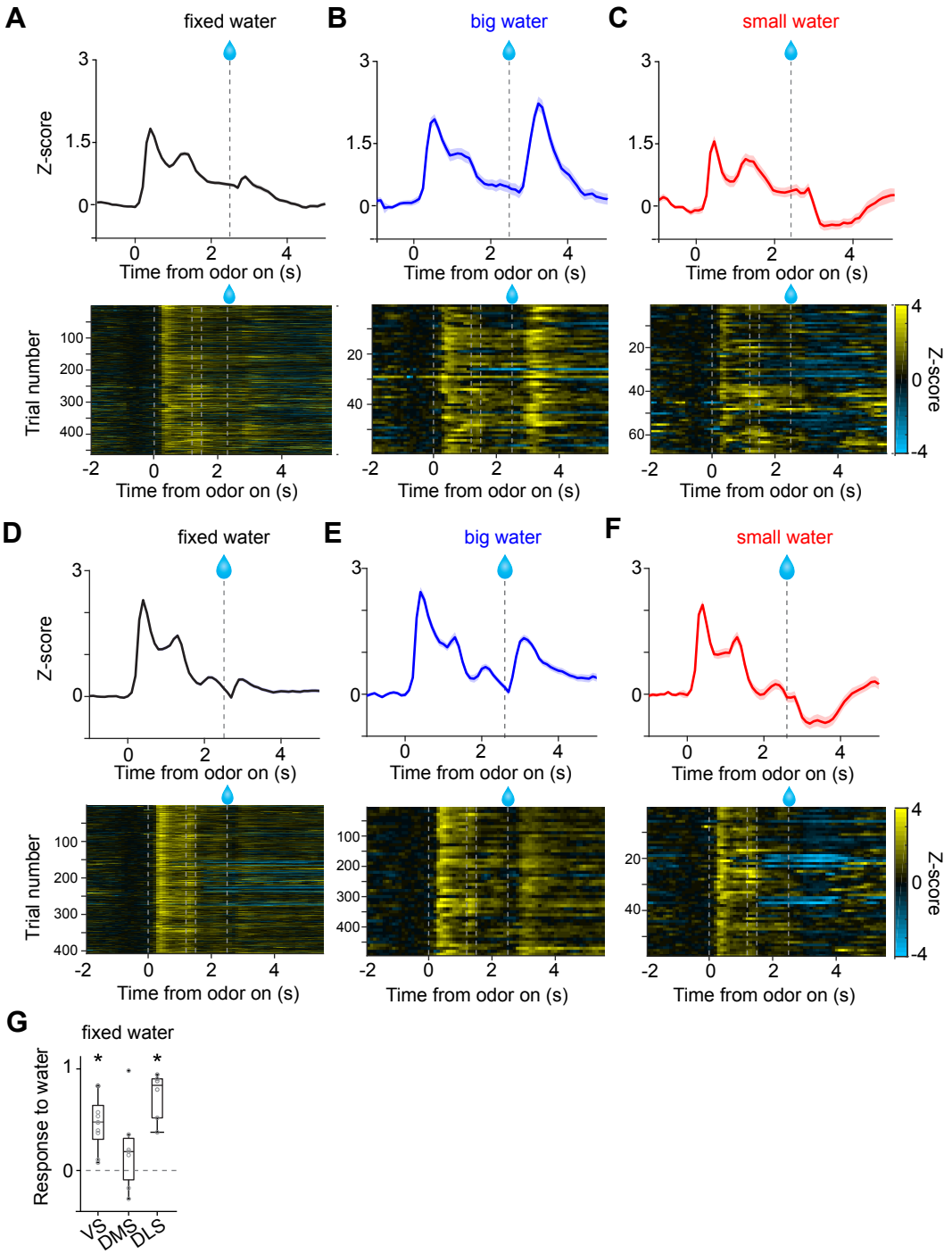
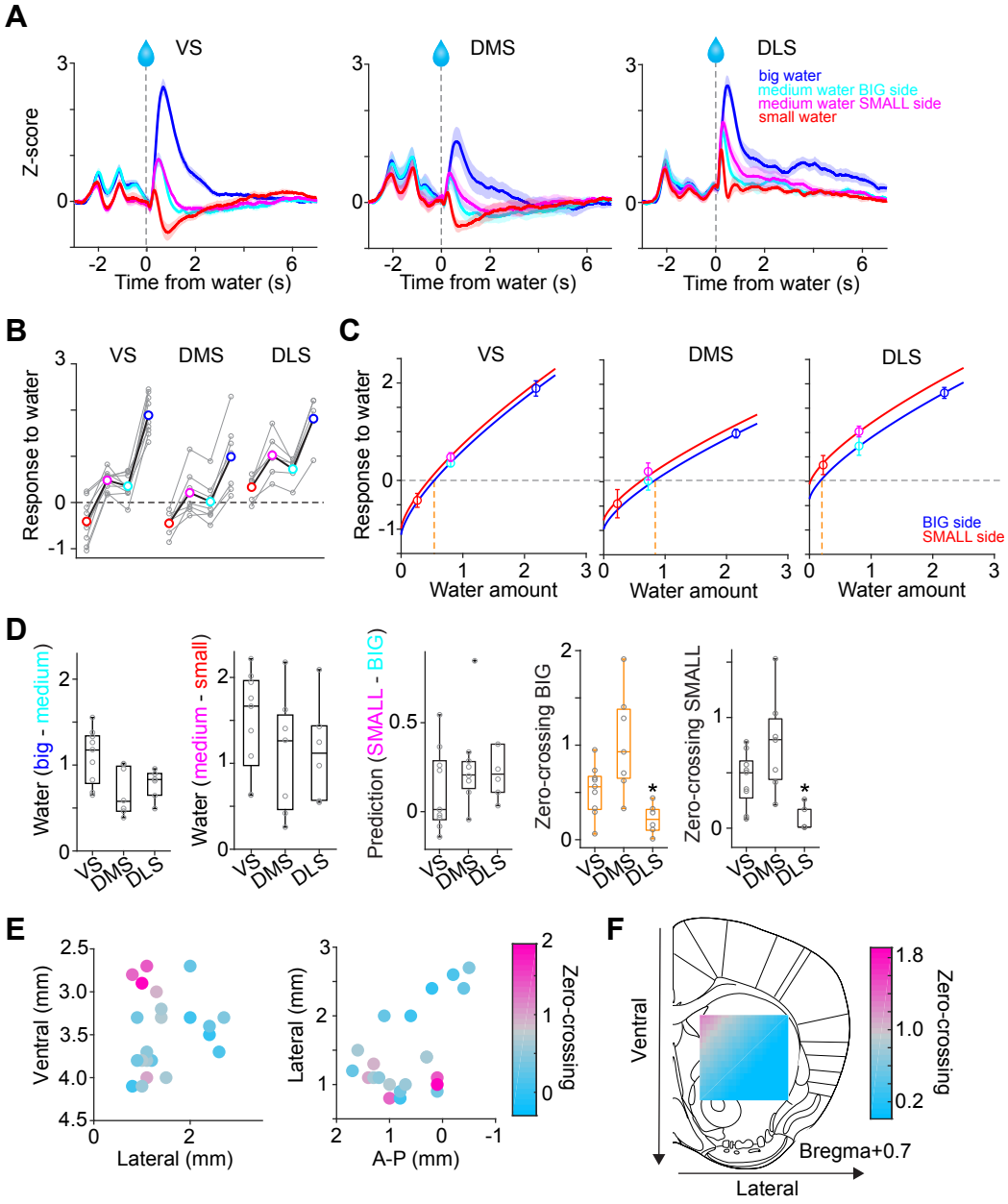


Figure 3. Small responses to fixed amounts of water in dopamine axons in DMS

(A, D) Dopamine axon responses to water in a fixed reward amount task (pure odor, correct choice). (B, E) Dopamine axon responses to a big amount of water in a variable reward amount task (pure odor, correct choice). (C, F) Dopamine axon responses to a small amount of water in a variable reward amount task (pure odor, correct choice). A-C, dopamine axon activity in an example animal; D-F, another example animal. (G) Responses to water (0.3-1.3 s after water onset) were significantly modulated with striatal location ($p=0.020$, ANOVA). The water responses were significantly positive in VS ($p=0.0011$) and in DLS ($p=6.3\times10^{-4}$), but not in DMS ($p=0.28$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4

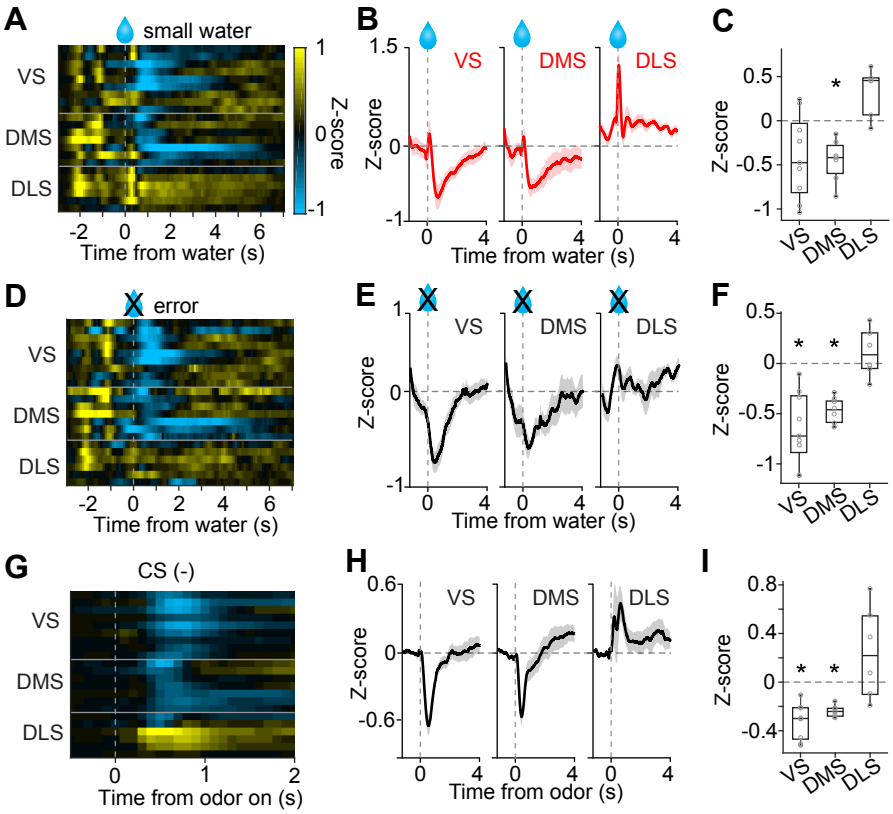


bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. Responses to water in dopamine axons in the striatum (A) Activity patterns per different striatal location, aligned at water onset (mean ± SEM, n=9 for VS, n=7 for DMS, n=6 for DLS). (B) Average responses to each water condition in each animal grouped by striatal areas. (C) Average response functions of dopamine axons in each striatal area. (D) Comparison of parameters for each animal grouped by striatal areas. "Water big-medium" is responses to big water minus responses to medium water at the BIG side and "Water medium-small" is responses to medium water minus responses to small water at the SMALL side, normalized with difference of water amounts (2.2 minus 0.8 for BIG and 0.8 minus 0.2 for SMALL). "Prediction SMALL-BIG" is responses to medium water at SMALL side minus responses to medium water at BIG side. "Zero-crossing BIG" is the water amount when the dopamine response is zero at BIG and side, which was estimated by the obtained response function. "Zero-crossing SMALL" is the water amount when the dopamine response is zero at SMALL side, which was estimated by the obtained response function. Response changes by water amounts (BIG or SMALL) or prediction was not significantly modulated by the striatal areas ($p=0.011$, $p=0.34$, $p=0.23$, ANOVA), whereas zero-crossing points (BIG or SMALL) were significantly modulated ($p=0.002$, $p=0.002$, ANOVA; $p=0.004$, DMS versus DLS for BIG side; $p=0.005$, VS versus DLS; $p=0.003$, DMS versus DLS for SMALL side). (E) Zero-crossing points were plotted along anatomical location in the striatum. Zero-crossing points were correlated with medial-lateral positions ($p=0.011$) and with dorsal-ventral positions ($p=0.014$). (F) Zero-crossing points were fitted with recorded location, and the estimated values in the striatal area were overlaid on the atlas for visualization (see Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5

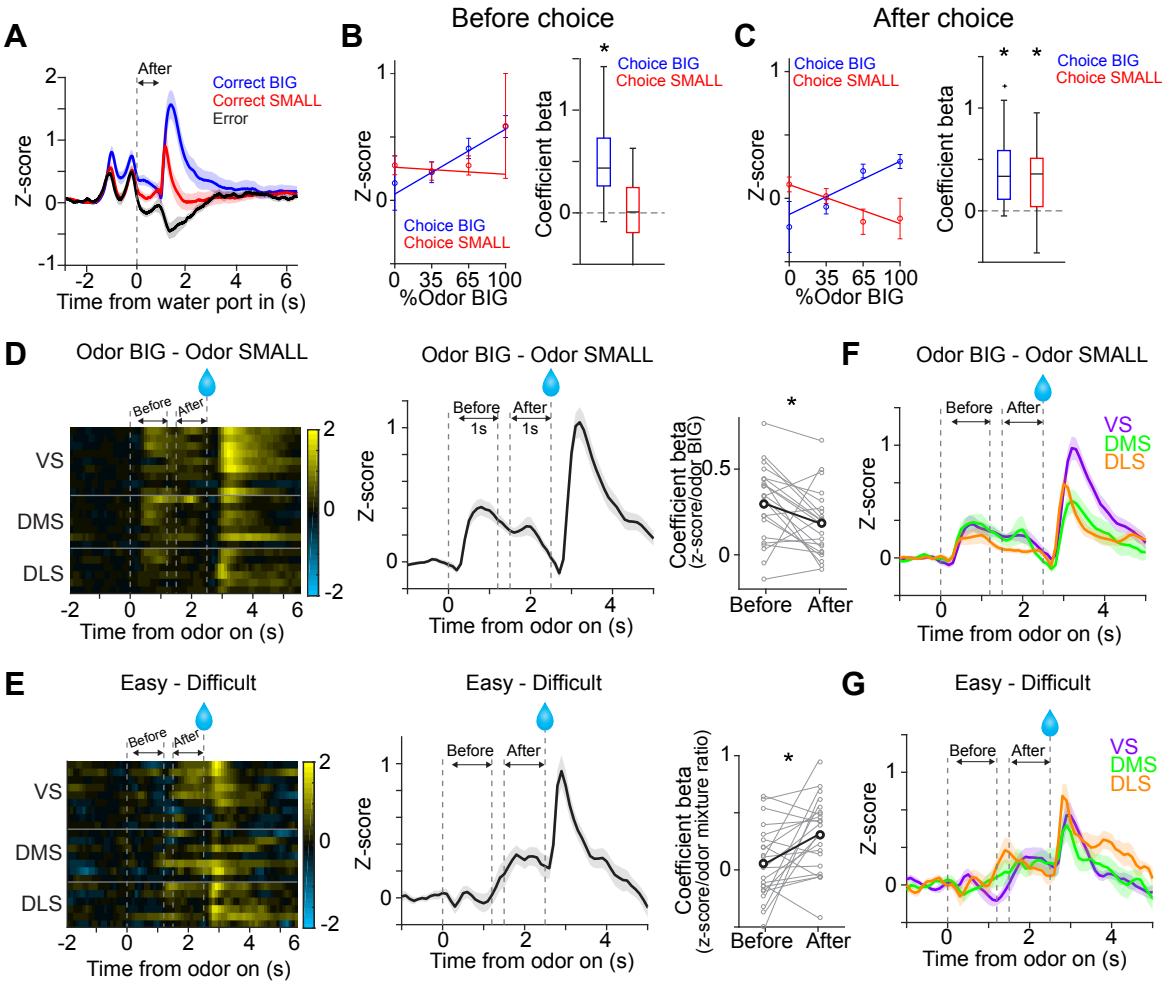


bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5. No inhibition by negative prediction error in dopamine axons in DLS (A) Activity pattern in each recording site aligned at small water. (B) Average activity pattern in each brain area (mean ± SEM). (C) Mean responses to small water (0.3-1.3 s after water onset) were negative in VS and DMS ($p=0.031$, $p=0.0025$, responses versus baseline), but not in DLS. The responses were different across striatal areas ($p=0.0013$, ANOVA; $p=0.0042$, VS versus DLS; $p=2.8\times10^{-4}$, DMS versus DLS). (D) Activity pattern aligned at water timing in error trials. (E) Average activity pattern in each brain areas (mean ± SEM). (F) Mean responses in error trials (0.3-1.3 s after water timing) were negative in VS and DMS ($p=6.2\times10^{-4}$, $p=6.5\times10^{-5}$, responses versus baseline), but not in DLS. The responses were different across striatal areas ($p=1.5\times10^{-4}$, ANOVA; $p=5.8\times10^{-4}$, VS versus DLS; $p=1.6\times10^{-4}$, DMS versus DLS). (G) Activity pattern aligned at CS(-) in a fixed reward amount task. (H) Average activity pattern in each brain area (mean ± SEM). (I) Mean responses at CS(-) (-1-0 s before odor port out) were negative in VS and DMS ($p=1.4\times10^{-4}$, VS; $p=1.0\times10^{-5}$, DMS, responses versus baseline), but not in DLS. Responses were different across striatal areas ($p=2.5\times10^{-4}$, ANOVA; $p=0.0012$, VS versus DLS; $p=0.0065$, DMS versus DLS).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6

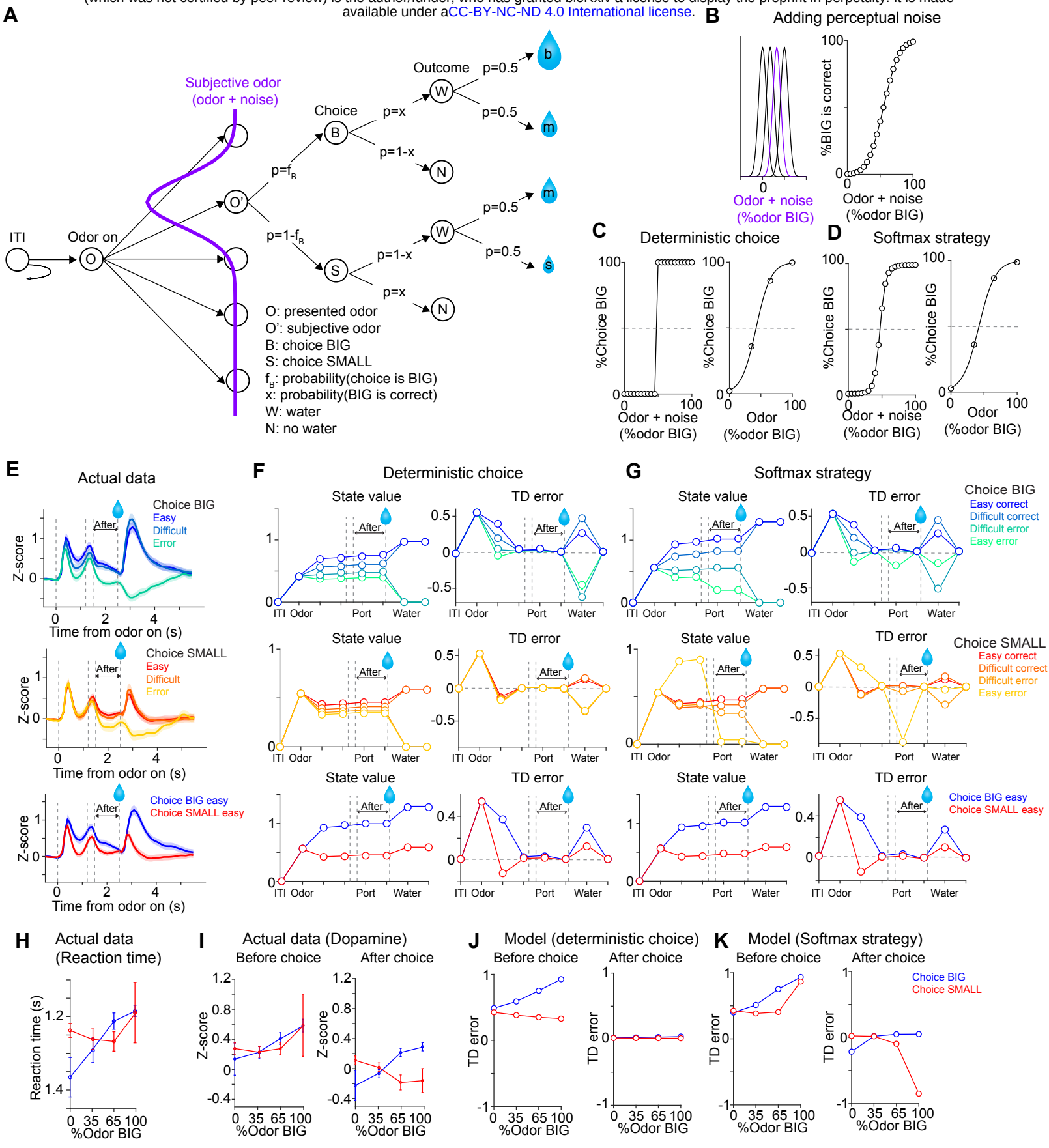


bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6. Dopamine signals stimulus-associated value and sensory evidence with different dynamics (A) Dopamine axon activity pattern aligned to time of water port entry for all animals (mean ± SEM). (B) Responses before choice (-1-0 s before odor port out) were fitted with linear regression with odor mixture ratio, and coefficient beta (slope) for all the animals are plotted. Correlation slopes were significantly positive for choice of the BIG side ($p=5.6\times10^{-6}$), but not significant for choice of the SMALL side ($p=0.42$). (C) Responses after choice (0-1 s after water port in) were fitted with linear regression with stimulus evidence (odor %) and coefficient beta (slope) for all the animals are plotted. Correlation slopes were significantly positive for both choice of the BIG side ($p=1.4\times10^{-5}$) and of the SMALL side ($p=2.2\times10^{-4}$). (D) Dopamine axon activity with an odor that instructed to choose BIG side (pure odor, correct choice) minus activity with odor that instructed to choose SMALL side (pure odor, correct choice) in each recording site (left), and the average difference in activity was plotted (mean ± SEM, middle). Correlation slopes between responses and stimulus-associated value (water amounts) significantly decreased after choice ($p=0.025$, before choice (-1-0 s before odor port out) versus after choice (0-1 s after water port in), pure odor, correct choice). (E) Dopamine axon activity when an animal chose SMALL side in easy trials (pure odor, correct choice) minus activity in difficult trials (mixture odor, wrong choice) in each recording site (left), and the average difference in activity was plotted (mean ± SEM, center). Coefficient beta between responses to odors and sensory evidence (odor %) significantly increased after choice ($p=0.0078$, before choice versus after choice). (F) Average difference in activity (odor BIG minus odor SMALL) before and after choice in each striatal area. The difference of coefficient (before versus after choice) was not significantly different across areas ($p=0.86$, ANOVA). (G) Average difference in activity (easy minus difficult) in each striatal area. The difference of coefficient (before versus after choice) was not significantly different across areas ($p=0.25$, ANOVA).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 7



bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 7. TD error dynamics capture emergence of sensory evidence after stimulus-associated value in dopamine axon activity (A) Trial structure in the model. Some repeated states are omitted for clarification. (B-D) Models were constructed by adding perceptual noise with normal distribution to each experimenter's odor (B left, subjective odor), calculating correct choice for each subjective odor (B right), and determining choice for each subjective odor (C or D left) according to choice strategy in the model. The final choice for each objective odor by experimenters (odor %) was calculated as the weighted sum of choice for subjective odors (C or D right). (E) Dopamine axon activity in trials with different levels of stimulus evidence: easy (pure odor, correct choice), difficult (mixture odor, correct choice), and error (mixture odor, error), when animals chose the BIG side (top) and when animals chose the SMALL side (middle). Bottom, dopamine axon activity when animals chose the BIG or SMALL side in easy trials (pure odor, correct choice). (F, G) Time-course in each trial of value (left) and TD error (right) of a model. (H) Line plots of actual reaction time from Figure 1G. Y-axis are flipped for better comparison with models. (I) Line plots of actual dopamine axon responses before and after choice from Figures 6B and 6C. (J, K) Model responses before and after choice were plotted with sensory evidence (odor %).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S1

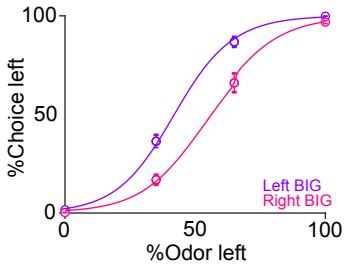


Figure S1. Average psychometric curve in odor manipulation blocks

% of choice of a left port when a left port is the BIG side or when a right port is the BIG side (mean ± SEM) and the average psychometric curve for each case.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S2

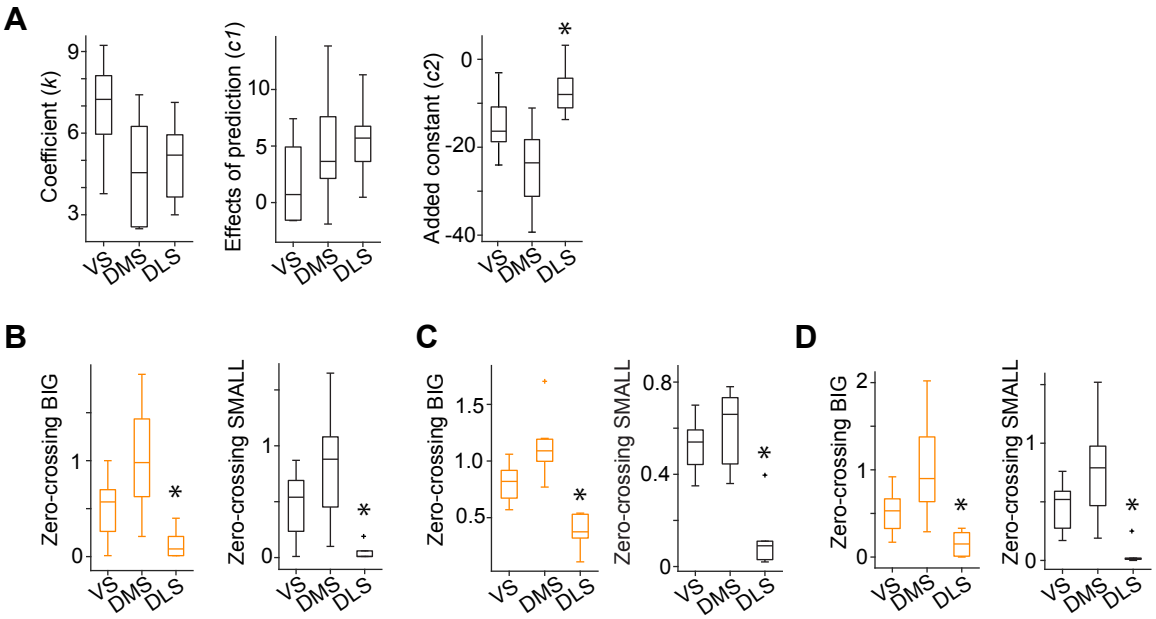


Figure S2. Zero-crossing points across the striatum with different methods (A) Each regression coefficient in the response function shown in Figure 4C. Fitting was performed by response $= k(R^{\alpha} + c1 \times S + c2)$, where R is the water amount, S is SMALL side (see Methods). (B) Zero-crossing points with linear function ($p=0.003$ for BIG; $p=6.1\times10^{-4}$ for SMALL, ANOVA). (C) Zero-crossing points with power function using a before-water time window (-1 to -0.2 s before water) as baseline. ($p=5.8\times10^{-5}$ for BIG; $p=2.1\times10^{-4}$ for SMALL, ANOVA). (D) Zero-crossing points using kernel models with power function ($p=0.0033$ and $p=8.9\times10^{-4}$, ANOVA).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S3

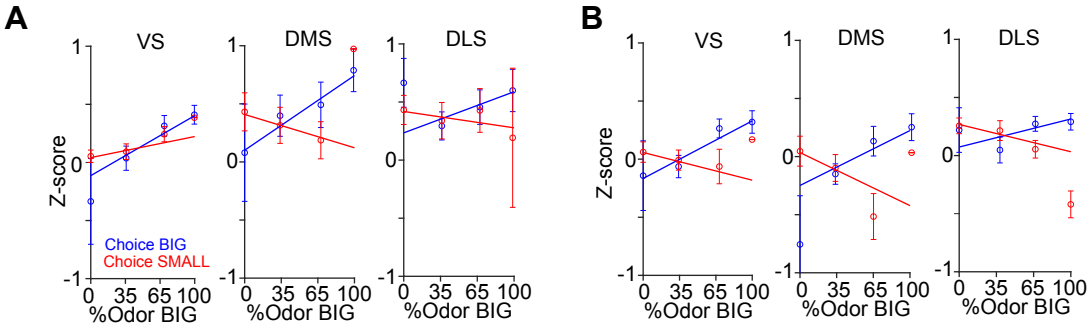


Figure S3. Dopamine axon responses before and after choice in each striatal area (A) Responses before choice (-1-0 s before odor port out) was fitted with linear regression with sensory evidence (odor %) and average fitted lines in each striatal area were plotted. The correlation slope for small choice was slightly modulated by striatal areas (p=0.0043, ANOVA; p=0.0013, VS versus DMS). (B) Responses after choice (0-1 s after water port in) was fitted with linear regression with sensory evidence and an average fitted line of each striatal area was plotted. The correlation slope was not significantly modulated by striatal areas (p=0.35 for choice BIG; p=0.35 for choice SMALL, ANOVA).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.22.262972; this version posted August 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

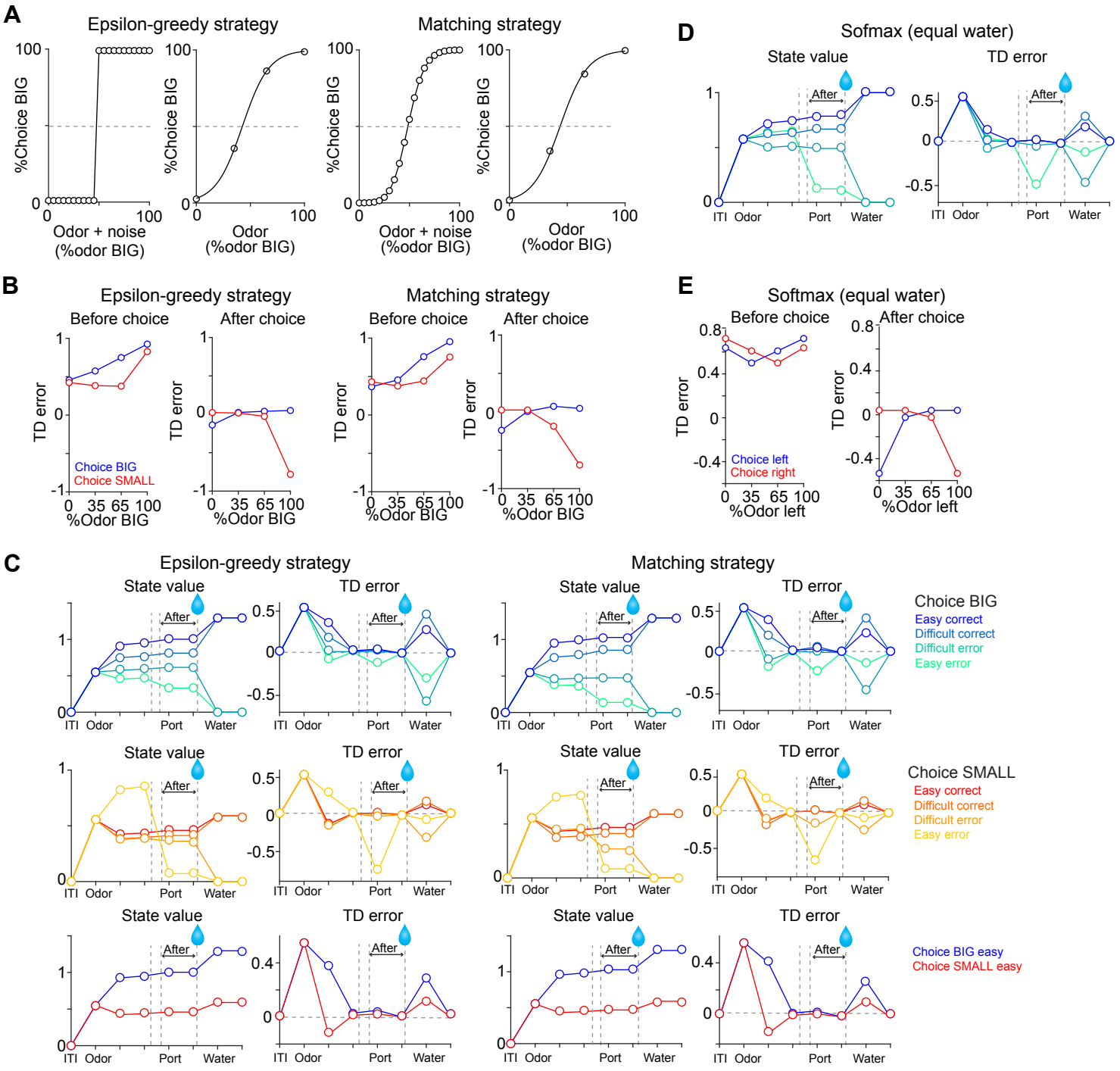


Figure S4. TD errors with stochastic choice strategies. (A) choice for each subjective odor (left) and choice for each objective odor (right) with epsilon greedy strategy and matching strategy. (B) TD errors with different sensory evidence (odor %) before and after choice in each model. (C) The temporal dynamics of state values and TD errors in each model. (D) The temporal dynamics of state values and TD errors with a softmax choice strategy (Figure 7D) but with equal amounts of water for both water ports. (E) TD errors with different levels of sensory evidence (odor %) before and after choice in model from D.