1 **Distinct temporal difference error signals in dopamine axons in three regions**

2 **of the striatum in a decision-making task**

3

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

5

6

7

8

9 **Affiliations:**

10 [1]Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University,

11 Cambridge, MA 02138, USA

12 [2]Department of Physiology, Osaka City University Graduate School of Medicine, Osaka, 545-8585, Japan

13 [3]Lead Contact

14

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

16

17

18

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

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

77 SNc (da Silva et al., 2018) as well as the axonal activity in the dorsal striatum (Howe and

78 Dombeck, 2016). Another study showed that dopamine axons in the dorsomedial striatum

79 (DMS) are activated when the animal makes a contralateral orienting movement in a decision-

80 making task (Parker et al., 2016). Other studies have also found that dopamine axons in the

81 posterior or ventromedial parts of the striatum are activated by aversive or threat-related stimuli

82 (de Jong et al., 2019; Menegas et al., 2017). An emerging view is that dopamine neurons

83 projecting to different parts of the striatum convey distinct signals and support different

84 functions (Cox and Witten, 2019).

85

86 Previous studies have shown that different parts of the striatum control distinct types of reward-

87 oriented behaviors (Dayan and Berridge, 2014; Graybiel, 2008; Malvaez and Wassum, 2018;

88 Rangel et al., 2008). First, the ventral striatum (VS) has often been associated with Pavlovian

89 behaviors, where the expectation of reward triggers relatively pre-programmed behaviors

90 (approaching, consummatory behaviors etc.) (Dayan and Berridge, 2014). Psychological studies

91 suggest that these behaviors are driven by stimulus-outcome associations (Kamin, 1969; Pearce

92 and Hall, 1980; Rescorla and Wagner, 1972). Consistent with this idea, previous experiments

93 have shown that dopamine in VS conveys canonical RPE signals (Menegas et al., 2017; Parker et

94 al., 2016), and support learning of values associated with specific stimuli (Clark et al., 2012). In

95 contrast, the dorsal part of the striatum has been linked to instrumental behaviors, where animals

96 acquire an arbitrary action that leads to a reward (Montague et al., 1996; Suri and Schultz, 1999).

97 Instrumental behaviors are further divided into two distinct types: goal-directed and habit

98 (Dickinson and Weiskrantz, 1985). Goal-directed behaviors are "flexible" reward-oriented

99 behaviors that are sensitive to a causal relationship ("contingency") between action and outcome,

100 and can quickly adapt to changes in the value of the outcome (Balleine and Dickinson, 1998).

101 After repetition of a goal-directed behavior, the behavior can become a habit which is

102 characterized by insensitivity to changes in the outcome value (e.g. devaluation) (Balleine and

103 O'Doherty, 2010). According to psychological theories, goal-directed and habitual behaviors are

104 supported by distinct internal representations: action-outcome and stimulus-response

105 associations, respectively (Balleine and O'Doherty, 2010). Lesion studies have indicated that

106 goal-directed behaviors and habit are controlled by DMS and the dorsolateral striatum (DLS),

107 respectively (Yin et al., 2004, 2005).

4

108

109 Instrumental behaviors are shaped by reward, and it is generally thought that dopamine is

110 involved in their acquisition (Gerfen and Surmeier, 2011; Montague et al., 1996; Schultz et al.,

111 1997). However, how dopamine is involved in distinct types of instrumental behaviors remain

112 unknown. A prevailing view in the field is that habit is controlled by "model-free" reinforcement

113 learning, while goal-directed behaviors are controlled by "model-based" mechanisms (Daw et

114 al., 2005; Dolan and Dayan, 2013; Rangel et al., 2008). In this framework, habitual behaviors are

115 driven by "cached" values associated with specific actions (action values) which animals learn

116 through direct experiences via dopamine RPEs. In contrast, goal-directed behaviors are

117 controlled by a "model-based" mechanism whereby action values are computed by mentally

118 simulating which sequence of actions lead to which outcome using a relatively abstract

119 representation (model) of the world. Model-based behaviors are more flexible compared to

120 model-free behaviors because a model-based mental simulation may allow the animal to

121 compute values in novel or changing circumstances. Although these ideas account for the

122 relative inflexibility of habit over model-based, goal-directed behaviors, they do not necessarily

123 explain the most fundamental property of habit, that is, its insensitivity to changes in outcome, as

124 cached values can still be sensitive to RPEs when the actual outcome violates expectation,

125 posing a fundamental limit in this framework (Dezfouli and Balleine, 2012; Miller et al., 2019).

126 Furthermore, the idea that habits are supported by action value representations does not

127 necessarily match with the long-held view of habit based on stimulus-response associations.

128

129 Until recently an implicit assumption across many studies was that dopamine neurons broadcast

130 the same teaching signals throughout the striatum to support different kinds of learning (Rangel

131 et al., 2008; Samejima and Doya, 2007). However, as mentioned before, more recent studies

132 revealed different dopamine signals across striatal regions, raising the possibility that different

133 striatal regions receive distinct teaching signals. In any case, few studies have directly examined

134 the nature of RPE signals across striatal regions in instrumental behaviors, in particular, between

135 DLS and other regions. As a result, it remains unclear whether different striatal regions receive

136 distinct dopamine signals during instrumental behaviors. Are dopamine signals in particular

137 areas dominated by movement-related signals? Are dopamine signals in these areas still

138 consistent with RPEs or are they fundamentally distinct? How are they different? Characterizing

139   dopamine signals in different regions is a critical step toward understanding how dopamine may

140   regulate distinct types of behavior.

141

142   In the present study, we sought to characterize dopamine signals in different striatal regions (VS,

143   DMS and DLS) during instrumental behaviors. We used a task involving both perceptual and

144   value-based decisions in freely-moving mice – a task that is similar to those previously used to

145   probe various important variables in the brain such as values, biases (Rorie et al., 2010; Wang et

146   al., 2013), confidence (Hirokawa et al., 2019; Kepecs et al., 2008), belief states (Lak et al.,

147   2017), and response vigor (Wang et al., 2013). In this task, the animal goes through various

148   movements and mental processes – self-initiating a trial, collecting sensory evidence, integrating

149   the sensory evidence with reward information, making a decision, initiating a choice movement,

150   committing to an option and waiting for reward, receiving an outcome of reward or no reward,

151   and adjusting internal representations for future performance using RPEs and confidence.

152   Compared to Pavlovian tasks, which have been more commonly used to examine dopamine

153   RPEs, the present task has various factors with which to contrast dopamine signals between

154   different areas.

155

156   Contrary to our initial hypothesis, dopamine signals in all three areas showed similar dynamics,

157   going up and down in a manner consistent with TD errors, reflecting moment-by-moment

158   *changes* in the expected future reward (i.e. state values). Notably, although we observed

159   correlates of accuracy and confidence in dopamine signals, consistent with previous studies

160   (Engelhard et al., 2019; Lak et al., 2017), the appearance of these variables was timing- and trial

161   type-specific. In stark contrast, our modeling demonstrate that these apparently diverse dopamine

162   signals can be inclusively explained by a single variable – TD error, that is moment-by-moment

163   changes in the expected reward in each trial. In addition, we found consistent differences

164   between these areas. For instance, DMS dopamine signals were modulated by contralateral

165   orienting movements, as reported previously (Parker et al., 2016). Furthermore, DLS dopamine

166   signals, while following TD error dynamics, were overall more positive, compared to other

167   regions. Based on these findings, we present novel models of how these distinct dopamine

168   signals may give rise to distinct types of behavior such as flexible versus habitual behaviors.

169

170

RESULTS

172

A perceptual decision-making task with reward amount manipulations

174

175 Mice were first trained in a perceptual decision-making task using olfactory stimuli (Figure 1)

176 (Uchida and Mainen, 2003). To vary the difficulty of discrimination, we used two odorants

177 mixed with different ratios (Figure 1A). Mice were required to initiate a trial by poking their

178 nose into the central odor port, which triggered a delivery of an odor mixture. Mice were then

179 required to move to the left or right water port depending on which odor was dominant in the

180 presented mixture. Odor-water side (left or right) rule was held constant throughout training and

181 recording in each animal. In order to minimize temporal overlaps between different trial events

182 and underlying brain processes, we introduced a minimum time required to stay in the odor port

183 (for 1 s before exiting the odor port) and in the water port (for 1 s) to receive a water reward.

184

185 After mice learned the task, the water amounts at the left and right water ports were manipulated

186 (Lak et al., 2017; Rorie et al., 2010; Wang et al., 2013) in a probabilistic manner. In our task, one

187 of the reward ports was associated with a big or medium size of water (BIG side) while another

188 side was associated with a small or medium size of water (SMALL side) (Figure 1A). In a daily

189 session, there were two blocks of trials, the first with equal-sized water and the second with

190 different distributions of water sizes on the two sides (BIG versus SMALL side). The reward

191 ports for BIG or SMALL conditions stayed unchanged within a session, and were randomly

192 chosen for each session. In each reward port (BIG or SMALL side), which of the two reward

193 sizes was delivered was randomly assigned in each trial. Note that the medium-sized reward is

194 delivered with the probability of 0.5 for every correct choice at either side. This design was used

195 to facilitate our ability to characterize RPE-related responses even after mice were well trained

196 (Tian et al., 2016). First, the responses to the medium sized-reward allowed us to characterize

197 how "reward expectation" affects dopamine reward responses because we can examine how

198 different levels of expectation, associated with the BIG and SMALL side, affect dopamine

199 responses to reward of the same (medium) amount. Conversely, for a given reward port, two

7

200   sizes of reward allow us to characterize the effect of "actual reward" on dopamine responses, by

201   comparing the responses when the actual reward was smaller versus larger than expected.

202

203   We first characterized the choice behavior by fitting a psychometric function (a logistic

204   function). Compared to the block with equal-sized water, the psychometric curve was shifted

205   laterally to the BIG side (Figure 1B, Figure S1). The fitted psychometric curves were laterally

206   shifted whereas the slopes were not significantly different across blocks (p=0.45) (Figure 1B).

207   We, therefore, quantified a choice bias as a lateral shift of the psychometric curve with a fixed

208   slope in terms of the % mixture of odors for each mouse (Figure 1C) (Wang et al., 2013). All the

209   mice exhibited a choice bias toward the BIG side (22/22 animals). Because a "correct" choice

210   (i.e. whether a reward is delivered or not) was determined solely by the stimulus in this task,

211   biasing their choices away from the 50/50 boundary inevitably lowers the choice accuracy (or

212   equivalently, the probability of reward). For ambiguous stimuli, however, mice could go for a

213   big reward, even sacrificing accuracy, in order to increase the long-term gain. Indeed, the

214   observed biases approximated the optimal bias that maximizes total reward (1.016 ± 0.001 times

215   reward compared to no bias, mean $\pm$ s.e.m, slightly less than the optimal bias that yields 1.022

216   times reward compared to no bias), rather than maximizing the accuracy (= reward probability,

217   i.e. no bias) or solely minimizing the risk (the variance of reward amounts) (Figures 1D and 1E).

218

219   Previous studies have shown that animals shift their decision boundary even without reward

220   amount manipulations in perceptual decision tasks (Lak et al., 2020a). These shifts occur on a

221   trial-by-trial basis, following a win-stay strategy, choosing the same side when that side was

222   associated with reward in the previous trial, particularly when the stimulus was more ambiguous

223   (Lak et al., 2020a). In the current task design, however, the optimal bias is primarily determined

224   by the sizes of reward (more specifically, which side delivered a big or small reward) which

225   stays constant across trials within a session. To determine whether the animal adopted a short-

226   time scale updating or a more stable bias, we next examined how receipt of reward affected the

227   choice in the subsequent trials. To extract trial-by-trial updating, we compared the psychometric

228   curves 1 trial before (n-1) and after (n+1) the current trials (n). This analysis was performed

229   separately for the rewarded side in the current (n) trials. We found that choice biases before and

230   after a specific reward location were not significantly different in any trial types (Figure 1F),

8

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

261 The constructed model captured modulations of dopamine axon activity time-locked to different

262 events (Figure 2C). On average, the magnitude of the extracted odor-locked activity was

263 modulated by odor cues. Dopamine axons were more excited by a pure odor associated with the

264 BIG side than a pure odor associated with the SMALL side (Figures 2C and 2F). The movement-

265 locked activity was stronger for a movement toward the contra-lateral, compared to the ipsi-

266 lateral side, which was most evident in DMS (Parker et al., 2016) but much smaller in VS or

267 DLS (Figure 2E, %Explained by movement). The choice-locked activity showed two types of

268 modulations (Figure 2C). First, it exhibited an inhibition in error trials at the time of reward (i.e.

269 when it has become clear that reward is not going to come). Second, dopamine activity showed a

270 modulation around the time of water port entry, an excitation when the choice was correct, and

271 an inhibition when the choice was incorrect, even before the mice received a feedback. These

272 "choice commitment"-related signals will be further analyzed below. Finally, delivery of water

273 caused a strong excitation which was modulated by the reward size (Figures 2C and 2F).

274 Furthermore, the responses to medium-sized water was slightly but significantly smaller on the

275 BIG side compared to the SMALL side (Figures 2C and 2F). The contribution of water-locked

276 kernels was larger than other kernels except in DMS, where odor, movement and water kernels

277 contributed similarly (Figures 2D and 2E).

278

279 In previous studies, RPE-related signals have typically been characterized by phasic responses to

280 reward-predictive cues and a delivery or omission of reward. Overall, the above results

281 demonstrate that observed populations contain the basic response characteristics of RPEs. First,

282 dopamine axons were excited by reward-predicting odor cues, and the magnitude of the response

283 was stronger for odors that instructed the animal to go to the side which is associated with a

284 higher value (i.e. BIG side). Responses to water were modulated by reward amounts, and the

285 water responses were suppressed by higher reward expectation. These characteristics were also

286 confirmed by using the actual responses, instead of obtained kernel models (Figures 2F and 2G).

287 Finally, in error trials, dopamine axons were inhibited when the time passed beyond the expected

288 time of reward, as the negative outcome becomes certain (Figure 2C). Next, we will investigate

289 each striatal area in more detail.

290

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

292

293   Although excitation to unpredicted reward is one of the signatures of dopamine RPE, recent

294   studies found that the dopamine axon response to water is small or undetectable in some part of

295   the dorsal striatum (Howe and Dombeck, 2016; Parker et al., 2016; da Silva et al., 2018).

296   Therefore, the above observation that all three areas (VS, DMS, and DLS) exhibited modulation

297   by reward may appear at odds with previous studies.

298

299   We noticed greatly diminished water responses when the reward amount was not manipulated,

300   that is, when dopamine axon signals were monitored during training sessions before introducing

301   the reward amount manipulations (Figure 3). In these sessions, dopamine axons in some animals

302   did not show significant excitation to water rewards (Figures 3A and 3D). This "lack" of reward

303   response was found in DMS, consistent with previous studies (Parker et al., 2016), but not in VS

304   or DLS (Figure 3G). Surprisingly, however, DMS dopamine axons in the same animals showed

305   clear excitation when reward amount manipulations were introduced, particularly strongly

306   responding to a big reward (Figures 3B and 3E). Indeed, the response patterns were qualitatively

307   similar across different striatal areas (Figure 4); the reward responses in all the areas were

308   modulated by reward size and expectation, although the whole responses seem to be shifted

309   higher in DLS, and lower in DMS (Figures 4A and 4B). These results indicate that a stochastic

310   nature of reward delivery in our task enhanced or "rescued" reward responses in dopamine axons

311   in DMS.

312

313   The above results emphasized the overall similarity of reward responses across areas, but some

314   important differences were also observed. Most notably, although a delivery of a small reward

315   caused an inhibition of dopamine axons below baseline in VS and DMS, the activity remained

316   non-negative in DLS. The overall responses tended to be higher in DLS.

317

318   In order to understand the diversity of dopamine responses to reward, we examined modulation

319   of dopamine activity by different parameters (Figure 4D). First, the effect of the amount of

320   "actual" reward was quantified by comparing responses to different amounts of water for a given

321   cue (i.e. the same expectation). The reward responses in all areas were modulated by reward

322   amounts, with a slightly higher modulation by water amounts in VS (Figure 4D Water big-

11

323   medium, Water medium-small). Next, the effect of expectation was quantified by comparing the

324   responses to the same amounts of water with prediction of different amounts. Effects of reward

325   size prediction were not significantly different across areas, although VS showed slightly less

326   modulation with more variability (Figure 4D, prediction SMALL-BIG).

327

328   Next, we sought to characterize these differences between areas in simpler terms by fitting

329   response curves (response functions). Previous studies that quantified responses of dopamine

330   neurons to varied amounts of reward under different levels of expectation indicated that their

331   reward responses can be approximated by a common function, with different levels of

332   expectation just shifting the resulting curves up and down while preserving the shape (Eshel et

333   al., 2016). We, therefore, fitted dopamine axon responses with a common response function (a

334   power or linear function) for each expectation level (i.e. separately for BIG and SMALL) while

335   fixing the shape of the function (i.e. the exponent of the power function or the slope of the linear

336   function were fixed, respectively) (Figure 4C, Figure S2A). The obtained response functions for

337   the three areas recapitulated the main difference between VS, DMS and DLS, as discussed

338   above. For one, the response curves of DLS are shifted overall upward. This can be characterized

339   by estimating the amount of water that does not elicit a change in dopamine responses from

340   baseline firing ("zero-crossing point" or reversal point). The zero-crossing points, obtained from

341   the fitted curves, were significantly lower in DLS (Figures 4C and 4D). The results were similar

342   regardless of whether the response function was a power (power function $\alpha < 1$) or a linear

343   function ($\alpha = 1$) (Figure S2B). Similar results were obtained using the aforementioned kernel

344   models in place of the actual activity (Figure S2D).

345

346   Since the recording locations varied across animals, we next examined the relationship between

347   recording locations and the zero-crossing points (Figures 4E and 4F). The zero-crossing points

348   varied both along the medial-lateral and the dorsal-ventral axes (linear regression coefficient; $\beta =$

349   -44 [zero-crossing point water amounts/mm], $p = 0.008$ for medial-lateral axis; $\beta = -52$, $p =$

350   0.011 for the dorsal-ventral axis). Examination of each animal confirmed that DMS showed

351   higher zero-crossing points (upper-left in Figure 4E left) whereas DLS showed lower zero-

352   crossing points (upper-right cluster in Figure 4E right).

353

12

354 We next examined whether the difference in zero-crossing points manifested specifically during

355 reward responses or whether it might be explained by recording artifacts; upward and downward

356 shifts in the response function can be caused by a difference in baseline activity before trial start

357 (odor onset), and/or lingering activity of pre-reward activity owing to the relatively slow

358 dynamics of the calcium signals (a combination of calcium concentration and the indicator). To

359 examine these possibilities, the same analysis was performed after subtracting the pre-reward

360 signals (Figure S2C). We observed similar or even bigger differences in zero-crossing points

361 ($p=2.2\times10^{-5}$, analysis of variance [ANOVA]). These results indicate that the elevated or

362 decreased responses, characterized by different zero-crossing points, was not due to a difference

363 in "baseline" but was related to the difference that manifests specifically in responses to reward.

364

365 Considerably small zero-crossing points in dopamine axons in DLS were not due to a poor

366 sensitivity to reward amounts nor a poor modulation by expected reward (Figure 4D). Different

367 zero-crossing points, i.e. shifts of the boundary between excitation and inhibition at reward,

368 suggest biased representation of TD error in dopamine axons across the striatum. In TD error

369 models, difference in zero-crossing points may affect not only water responses but also responses

370 to other events. Thus, the small zero-crossing points in dopamine axons in DLS should yield

371 almost no inhibition following an event that is worse than predicted. To test this possibility, we

372 examined responses to events with lower value than predicted (Figure 5): small water (Figures

373 5A-5C), water omission caused by choice error (Figures 5D-5F), and a cue that was associated

374 with no outcome (Figures 5G-5I). Consistent with our interpretation of small zero-crossing

375 points, dopamine axons in DLS did not show inhibition in response to outcomes that were worse

376 than predicted while being informative about water amounts.

377

378 Taken together, these results demonstrate that dopamine reward responses in all three areas

379 exhibited characteristics of RPEs. However, relative to canonical responses in VS, the responses

380 were shifted more positively in the DLS and more negatively in the DMS.

381

382

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

384

385   The analyses presented so far mainly focused on phasic dopamine responses time-locked to cues

386   and reward. However, dopamine axon activity also exhibited richer dynamics between these

387   events, which need to be explained. For instance, the signals diverged between correct and error

388   trials even before the actual outcome was revealed (a reward delivery versus a lack thereof)

389   (Figure 2C Choice). This difference between correct and error trials, which is dependent on the

390   strength of sensory evidence (or stimulus discriminability), was used to study how neuronal

391   responses are shaped by "confidence". Confidence is defined as the observer's posterior

392   probability that their decision is correct given their subjective evidence and their choice

393   ($P(reward|stimulus, choice)$) (Hangya et al., 2016). A decision model allows the

394   experimenter to link stimulus discriminability to subjective evidence (Hangya et al., 2016). A

395   given model and task structure makes specific predictions on the shape of three key signatures

396   relating stimulus discriminability, choice and confidence. The predictions can vary depending on

397   task design (Adler and Ma, 2018; Rausch and Zehetleitner, 2019), but the structure of our task

398   follows the original predictions (Hangya et al., 2016). Additionally, in our task, the mice

399   combined the information about reward size with the strength of sensory evidence to select an

400   action (confidence, or uncertainty) (Figure 1). The previous analyses did not address how these

401   different types of information affect dopamine activity over time. We next sought to examine the

402   time course of dopamine axon activity in greater detail, and to determine whether a simple model

403   could explain these dynamics.

404

405   Our task design included two delay periods, imposed before choice movement and water

406   delivery, to improve our ability to separate neuronal activity associated with different processes

407   (Figure 1A). The presence of stationary moments before and after the actual choice movement

408   allows us to separate time windows before and after the animal's commitment to a certain option.

409   We examined how the activity of dopamine neurons changed before choice movement and after

410   the choice commitment (Figure 6).

411

412   We first examined dopamine axon activity after water port entry (0-1 s after water port entry). In

413   this period, the animals have committed to a choice and are waiting for the outcome to be

414   revealed. Responses following different odor cues were plotted separately for trials in which the

415   animal chose the BIG or SMALL side. The vevaiometric curve (a plot of responses against

14

416   sensory evidence) followed the expected 'X-pattern' with a modulation by reward size

417   (Hirokawa et al., 2019), which matches the expected value for these trial types, or the size of

418   reward multiplied by the probability of receiving a reward, given the presented stimulus and

419   choice (Figure 6C). The latter has been interpreted as the decision confidence,

420    $P(reward|stimulus, choice)$ (Lak et al., 2017, 2020b). The crossing point of the two lines

421   forming an "X" is shifted to the left in our data because of the difference in the reward size

422   (Figure 6C).

423

424   When this analysis was applied to the time period before choice movement (0-1 s before odor

425   port exit), the pattern was not as clear; the activity was monotonically modulated by the strength

426   of sensory evidence (%Odor BIG) only for the BIG choice trials, but not for the SMALL choice

427   trials (Figure 6B). This result is contrary to a previous study that suggested that the dopamine

428   activity reflecting confidence develops even before a choice is made (Lak et al., 2017). We note,

429   however, that the previous study only examined the BIG choice trials, and the results were

430   shown by "folding" the x-axis, that is, by plotting the activity as a function of the stimulus

431   contrast (which would correspond to |%Odor BIG – 50| in our task), with the result matching the

432   so-called "folded X-pattern". We would have gotten the same result, had we plotted our results

433   in the same manner excluding the SMALL choice trials. Our results, however, indicate that a full

434   representation of "confidence" only becomes clear after a choice commitment, leaving open the

435   question what the pre-choice dopamine activity really represents.

436

437   The aforementioned analyses, using either the kernel regression or actual activity showed that

438   cue responses were modulated by whether the cue instructed a choice toward the BIG or SMALL

439   side (Figures 2C and 2F). These results indicate that the information about stimulus-associated

440   values (BIG versus SMALL) affected dopamine neurons earlier than the strength of sensory

441   evidence (or confidence). We next examined the time course of how these two variables affected

442   dopamine axon activity more closely. We computed the dopamine axon activity between trials

443   when a pure odor instructed to go to the BIG versus SMALL side. Consistent with the above

444   result, the difference was evident during the cue period, and then gradually decreased after

445   choice movement (Figure 6D). We performed a similar analysis, contrasting between easy and

446   difficult trials (i.e. the strength of sensory evidence). We computed the difference between

15

447   dopamine axon activity in trials when the animal chose the SMALL side after the strongest

448   versus weaker stimulus evidence (a pure odor that instruct to choose the SMALL side versus an

449   odor mixture that instruct to choose the BIG side). In stark contrast to the modulation by the

450   stimulus-associated value (BIG versus SMALL), the modulation by the strength of stimulus

451   evidence in SMALL trials fully developed only after a choice commitment (i.e. water port entry)

452   (Figure 6E). Across striatal regions, the magnitude and the dynamics of modulation due to

453   stimulus-associated values and the strength of sensory evidence were similar (Figures 6F and

454   6G), although we noticed that dopamine axons in DMS showed slightly higher correlation with

455   sensory evidence before choice (Figure S3).

456

457   As discussed above, a neural correlate of "confidence" appears at a specific time point (after

458   choice commitment and before reward delivery) or in a specific trial type (when an animal would

459   choose BIG side) before choice. We, therefore, next examined whether a simple model can

460   account for dopamine axon activity more inclusively (Figure 7). To examine how the value and

461   RPE may change within a trial, we employed a Monte-Carlo approach to simulate animal's

462   choices assuming that the animal has already learned the task. We used a Monte-Carlo method to

463   obtain the ground truth landscape of the state values over different task states, without assuming

464   a specific learning algorithm.

465

466   The variability and errors in choice in psychophysical performance are thought to originate in the

467   variability in the process of estimating sensory inputs (perceptual noise) or in the process of

468   selecting an action (decision noise). We first considered a simple case where the model contains

469   only perceptual noise (Green and Swets, 1966). In this model, an internal estimate of the

470   stimulus or a "subjective odor" was obtained by adding Gaussian noise to the presented odor

471   stimulus on a trial-by-trial basis (Figures 7B left). In each trial, the subject chooses

472   deterministically the better option (Figure 7C left) based on the subjective odor and the reward

473   amount associated with each choice (Figure 7B right). The model had different "states"

474   considering N subjective odors (N = 60 and 4 were used and yielded similar results), the

475   available options (left versus right), and a sequence of task events (detection of odor, recognition

476   of odor identity, choice movement, water port entry [choice commitment], Water/No-water

477   feedback, inter-trial interval [ITI]) (Figure 7A). The number of available choices is two after

16

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

17

509   exploration of different options, and can be modeled using different methods or rationale behind

510   it. Here we present results based on a "softmax" decision rule, in which a decision variable (in

511   this case, the difference in the ratio of the expected values at the two options) was transformed

512   into the probability of choosing a given option using a sigmoidal function (e.g. Boltzmann

513   distribution) (Sutton and Barto, 2011). We also tested other stochastic decision rules such as

514   Herrnstein's matching law (Herrnstein, 1961) or ε-greedy exploration (randomly selecting an

515   action in a certain fraction [ε] of trials) (Sutton and Barto, 2011) (Figures S4A-S4C).

516

517   Interestingly, just by adding some stochasticity in action selection, various peculiar features of

518   dopamine axon signals described above were suddenly explained (Figures 7G and 7K). Note that

519   the main free parameters of the above models are the width of the Gaussian noise, which

520   determines the "slope" of the psychometric curve, and was chosen based merely on the

521   behavioral performance, but not the neural data. When the model chose the BIG side, state value

522   at odor presentation was roughly monotonically modulated by the strength of sensory evidence

523   similar to the above (Figure 7G top left). When the model chose the SMALL side, however, the

524   relationship between the strength of sensory evidence and value was more compromised (Figure

525   7G middle left). As a result, TD error did not show monotonic relationship with sensory

526   evidence before choice (Figures 7G middle right and 7K left), similar to actual dopamine axons

527   responses (Figures 7E middle and 7I left), which was reminiscent of reaction time pattern

528   (Figure 7H). On the other hand, once a choice was committed, the model exhibited interesting

529   dynamics very different from the above deterministic model. After choice commitment, expected

530   value was monotonically modulated by the strength of sensory evidence both for the choice to

531   the BIG and SMALL sides (Figure 7G top and middle left, After). Further, because of the

532   introduced stochasticity in action selection, the model sometimes chose a suboptimal option,

533   resulting in a drop in the state value. This, in turn, caused TD error to exhibit an "inhibitory dip"

534   once the model "lost" a better option (Figure 7G right), similar to the actual data (Figures 7E and

535   7I). This effect was strong particularly when the subjective odor instructed the BIG side but the

536   model ended up choosing the SMALL side. For a similar reason, TD error showed a slight

537   excitation when the model chose a better option (i.e. lost a worse option). The observed features

538   in TD dynamics were not dependent on exact choice strategy: softmax, matching, and ε-greedy,

539   all produced similar results (Figures S4B and S4C). This is because, with any strategy, after

540   commitment of choice, the model loses another option with a different value, which results in a

541   change in state value. These results are in stark contrast to the first model in which all the choice

542   errors were attributed to perceptual noise.

543

544   In summary, we found that a standard TD error, computing the moment-by-moment changes in

545   state value (or, the expected future reward), can capture various aspects of dynamics in dopamine

546   axon activity observed in the data, including the changes that occur before and after choice

547   commitment, and the detailed pattern of cue-evoked responses. These results were obtained as

548   long as we introduced some stochasticity in action selection (decision noise), regardless of how

549   we did it. The state value dynamically changes during the performance of the task because the

550   expected value changes according to an odor cue (i.e. strength of sensory evidence and stimulus-

551   associated values) and the changes in potential choice options. A drop of the state value and TD

552   error at the time of choice commitment occurs merely because the state value drops when the

553   model chose an option that was more likely to be an error. Further, a correlate of "confidence"

554   appears after committing a choice, merely because at that point (and *only* at that point), the state

555   value becomes equivalent to the reward size multiplied with the confidence, i.e. the probability

556   of reward given the stimulus and the choice. This means that, as long as the animal has

557   appropriate representations of states, a representation of "confidence" can be acquired through a

558   simple associative process or model-free reinforcement learning without assuming other

559   cognitive abilities such as belief states or self-monitoring (meta-cognition). In total, not only the

560   phasic responses but also some of the previously unexplained dynamic changes can be explained

561   by TD errors computed over the state value, provided that the model contains some stochasticity

562   in action selection in addition to perceptual noise. Similar dynamics across striatal areas (Figure

563   6) further support the idea that dopamine axon activity follows TD error of state values in spite

564   of the aforementioned diversity in dopamine signals.

565

19

566   **DISCUSSION**

567

568   In the present study, we monitored dopamine axon activity in three regions of the striatum (VS,

569   DMS and DLS) while mice performed instrumental behaviors involving perceptual and value-

570   based decisions. In addition to phasic responses associated with reward-predictive cues and

571   reward, we also analyzed more detailed temporal dynamics of the activity within a trial. Contrary

572   to the current emphases on diversity or multiplexing in dopamine signals (and therefore, to our

573   surprise), we found that dopamine axon activity in all of the three areas exhibited dynamics that

574   can be explained by the TD error which calculates moment-by-moment "changes" in the

575   expected future reward (i.e. state value). Interestingly, however, our results showed consistent

576   differences between regions. First, as reported previously (Parker et al., 2016), during choice

577   movements, contra-lateral orienting movements caused a transient activation in the DMS. This

578   response was negligible in VS and DLS, however. Second, although dopamine axon signals

579   exhibited temporal dynamics that are predicted by TD errors, reward responses were generally

580   elevated in DLS. As a consequence, dopamine axon signals in DLS did not exhibit a clear

581   inhibitory response ("dopamine dip") even when the actual reward was smaller than expected, or

582   even when the animal did not receive a reward, despite our observations that dopamine axons in

583   VS and DMS exhibited clear inhibitory responses in these conditions. Overall, the activity during

584   the reward period was biased toward positive responses in the DLS, compared to other areas.

585   Activation of dopamine neurons both in VTA and SNc are known to reinforce preceding

586   behaviors (Ilango et al., 2014; Keiflin et al., 2019; Lee et al., 2020; Saunders et al., 2018). The

587   differences in dopamine axon signals that we observed in instrumental behaviors can provide

588   specific constraints on the behaviors learned through dopamine-mediated reinforcement in these

589   striatal regions.

590

591   **Diversity in representation of TD errors**

592

593   Accumulating evidence indicates that dopamine neurons are diverse in various aspects such as

594   anatomy, physiological properties, and activity (Engelhard et al., 2019; Farassat et al., 2019;

595   Howe and Dombeck, 2016; Kim et al., 2015; Lammel et al., 2008; Matsumoto and Hikosaka,

596   2009; Menegas et al., 2015, 2017, 2018; Parker et al., 2016; da Silva et al., 2018; Watabe-Uchida

597   and Uchida, 2018). Our study is one of the first to examine dopamine signals in three different

598   regions of the striatum during an instrumental behavior involving perceptual and value-based

599   decisions. We found that dopamine axon activity in the striatum follows TD error dynamics in

600   our choice paradigm. At the same time, we found that the response function for water delivery in

601   dopamine axons in different striatal areas showed different zero-crossing points, the boundary

602   between excitatory and inhibitory responses (Figure 4). The results suggested that dopamine

603   axons in DMS use a higher boundary (requiring larger amounts of reward to excite), and

604   dopamine axons in DLS use a lower boundary (requiring smaller amounts of reward to excite).

605   In other words, dopamine signals in DMS use a strict criterium to be excited, whereas dopamine

606   signals in DLS tend to be more excited with smaller rewards.

607

608   A recent study (Dabney et al., 2020) proposed that the diversity in dopamine responses

609   potentially give rise to a population code for a reward distribution (distributional reinforcement

610   learning). In this theory, there are optimistic and pessimistic dopamine neurons. Optimistic

611   dopamine neurons emphasize positive over negative RPEs, and as a consequence, their

612   corresponding value predictors are biased to predict a higher value in a reward distribution, or

613   vice versa. The distributional reinforcement learning, as formulated in Dabney et al. (Dabney et

614   al., 2020), predicts that optimistic and pessimistic dopamine neurons should have zero-crossing

615   points shifted toward larger and smaller rewards, respectively. In this sense, our observation that

616   DLS dopamine signals have smaller zero-crossing points resembles pessimistic dopamine

617   neurons in distributional reinforcement learning, although the previous study found both

618   optimistic and pessimistic dopamine neurons in the VTA, which does not necessarily project to

619   the DLS. Whether the present result is related to distributional reinforcement learning requires

620   more specific tests such as dopamine neurons' sensitivity to positive versus negative RPEs

621   (Dabney et al., 2020). It will be interesting to characterize these response properties in a

622   projection-specific manner.

623

624   Higher criteria in DMS may partly explain the observation that some dopamine neurons do not

625   show a clear excitation by reward, such as in the case of our recording without reward amount

626   modulations (Figure 3). Our results suggest that whether dopamine neurons respond to reward

627   likely depends critically on task structures and training history. It will be important to further

21

628   examine in what conditions these dopamine neurons lose responses to water, or whether there are

629   dopamine neurons which do not respond to reward in any circumstances. In contrast to DMS, we

630   observed reliable excitation to water reward in dopamine axons in DLS. Thus, the previous

631   observation that some dopamine neurons in the substantia nigra show small or no excitation to

632   reward (da Silva et al., 2018) may mainly come from DMS-projecting dopamine neurons or

633   another subpopulation of dopamine neurons that project to the tail of the striatum (TS) (Menegas

634   et al., 2018), but not DLS. The distinction is important because smaller dopamine responses to

635   reward have been often linked to skill or habit with value-free mechanism (Miller et al., 2019).

636   In contrast, we found that dopamine axons in DLS show strong modulation by reward amounts

637   and prediction, and its dynamics resemble TD errors. Our observation suggests that the lack of

638   reward omission responses and excitation by even small rewards is a key for the function of

639   dopamine in DLS.

640

641

642   **Positively biased reinforcement signals in DLS dopamine**

643

644   It has long been observed that the activity of many dopamine neurons exhibits a phasic inhibition

645   when an expected reward was omitted or when the reward received was smaller than expected

646   (Hart et al., 2014; Schultz et al., 1997). This inhibitory response to negative RPEs is one of the

647   hallmarks of dopamine RPE signals. Our results that DLS dopamine signals largely lack these

648   dopamine dips (Figure 4 and Figure 5) has profound implications on what types of behaviors are

649   learned through DLS dopamine signals as well as what computational principles underlie

650   reinforcement learning in DLS.

651

652   Dopamine "dips" are thought to act as aversive stimuli and/or can facilitate extinction of

653   previously learned behaviors (weakening) (Chang et al., 2018; Montague et al., 1996; Schultz et

654   al., 1997). The lack of dopamine dip in DLS may lead to the animal's reduced sensitivity to

655   worse-than-expected outcome (i.e. negative prediction error). This characteristic resembles the

656   activity of dopamine axons in TS, posterior to DLS, which signals potential threat and also lacks

657   inhibitory responses to an omission of a predicted threat (Menegas et al., 2017, 2018). We

658   proposed that the lack of inhibitory omission signals (and so lack of weakening signals) would

659   be critical to maintain threat prediction even if an actual threat is sometimes omitted. Similarly,

660   the lack of weakening signals in DLS may help keep the learned actions from being erased even

661   if the outcome is sometimes worse than predicted or even omitted. This idea is in line with the

662   previous observations that DLS plays an important role in habitual behaviors (Yin et al., 2004).

663   The uniquely modified TD error signal in DLS (i.e. a reduced inhibitory response during the

664   reward period) may explain a predominant role of DLS in controlling habitual behaviors.

665

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

667

668   A deeper understanding of the nature of reinforcement signals can constrain the search for

669   computational principles and provide critical insight into what is actually learned by the system.

670   Here we speculate on these questions in the light of reinforcement learning theories and

671   anatomy.

672

673   Thorndike (Thorndike, 1932) proposed three principles for instrumental learning – the law of

674   effect, the law of readiness, and the law of exercise. The law of effect emphasizes the role of

675   outcome of behaviors: behaviors that led to good outcomes become more likely to occur – the

676   idea that is a foundation of value-based reinforcement learning. In contrast, the law of exercise

677   emphasizes the number of times a particular action was taken. There has been an increasing

678   appreciation of the law of exercise because repetition or overtraining is the hallmark of habits

679   and skills (Hikosaka et al., 1995; Matsuzaka et al., 2007; Miller et al., 2019; Morris and

680   Cushman, 2019; Ölveczky, 2011; Robbins and Costa, 2017; Smith and Graybiel, 2016). Here we

681   propose that dopamine signals in DLS provide an ideal neural substrate of learning with an

682   emphasis on the law of exercise. A positively biased TD error signals ensures that an "OK"

683   action will be positively reinforced, in a manner that depends on the number of times that the

684   same behavior was repeated as far as it is accompanied by a small reward (i.e. with "OK"

685   signals). This property may explain why the formation of habit (and skills) normally requires

686   overtraining (i.e. repeating a certain behavior many times).

687

688   The observation that DLS dopamine signals lack inhibitory responses raises the question what is

689   actually learned by the system. Learning of values depends on the balance between positive and

690   negative prediction errors: the learned value converges to the point at which positive and

691   negative prediction errors form an equilibrium. If a reinforcement signal lacks negative

692   prediction errors, this learning would no longer work as it was originally conceptualized. In

693   reinforcement learning theories, an alternative approach is policy-based reinforcement learning.

694   We propose that policy learning may be a better way to conceptualize the function of the DLS. In

695   reinforcement learning, a policy is a set of rules that map an action to a state, and has direct

696   relevance to stimulus-response associations that are proposed to underlie habit because the

697   relationship between stimulus (state) and response (action) can be more directly encoded using a

698   policy. According to Sutton and Barto (Sutton and Barto, 2018), policy learning can be done by

699   learning what is called "preference", $h(s, a)$, which defines the likelihood of a certain action, a,

700   in a given state, s. In a given state, an action is selected based on preference through a winner-

701   take-all mechanism either deterministically (e.g. by selecting the action with the maximum

702   preference) or stochastically (e.g. through a softmax action selection). One way to conceptualize

703   preference is to see it as a generalized version of value, which has less constraints than value (the

704   idea of "value" may imply many properties that it should follow, e.g. the value should be zero for

705   no outcome). Alternatively, $h(s, a)$ can directly encode the probability of an action.

706

707   It is also important to consider what are "states" for learning in DLS. Importantly, the main

708   inputs to DLS come from the motor cortex, somatosensory cortex, and other subcortical areas

709   such as intralaminar nuclei in thalamus (Hunnicutt et al., 2016). Thus, the inputs to DLS may not

710   be dominated by the sensory information representing the external world, as often

711   conceptualized in reinforcement learning. Instead, DLS is well-positioned to receiving inputs

712   representing motor commands (the current "motor states") or somatosensory information (the

713   current "bodily states" consisting of proprioception, sense of touch etc.). In other words, DLS

714   may compute their output by monitoring the current motor and bodily states. Dopamine in DLS

715   can thus be conceptualized as a reinforcement signal that strengthens the connection between the

716   current motor/bodily state and the next motor output. This mechanism, when chained, can

717   produce a sequence of movements as long as the same motor/bodily state is revisited or

718   reproduced, which may not occur easily at the beginning but can occur after repeated training.

719   As such, DLS may regulate "how" to perform a sequence of well-trained movements smoothly

720   and automatically. The key properties of habits and skills such as stereotypy, automaticity and

24

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

25

752   pushed up or down, whereas the information is still almost intact. In addition to anatomical

753   reasons, DLS-projecting dopamine neurons show higher burstiness in intact animals (Farassat et

754   al., 2019) and higher excitability *in vitro* (Evans et al., 2017; Lerner et al., 2015). These multiple

755   reasons may explain why DLS-projecting dopamine neurons do not show inhibitory responses to

756   negative prediction errors. It will be fascinating if we could connect all these levels of studies

757   into functional meaning in the future.

758

759   **Future directions to understand the meaning of diversity of dopamine signals**

760

761   Recent studies reported that dopamine neurons are modulated by various parameters (Engelhard

762   et al., 2019; Watabe-Uchida and Uchida, 2018). Here, we found that TD error dynamics can

763   inclusively explain two seemingly separate decision variables, namely, stimulus-associated value

764   and choice accuracy when animal's choice strategy is not deterministic (i.e. there is decision

765   noise) (Figure 6). At a glance, dopamine activity patterns may appear to be signaling two distinct

766   variables at different timings, but both are inclusively explained by a single quantity (TD error)

767   in one framework (Figure 7). These results underscore the importance of considering moment-

768   by-moment dynamics, and underlying computation. Taken together, our results showed that

769   dopamine axon signals in the striatum approximate TD error dynamics. We propose that

770   dopamine in different striatal areas conveys TD errors in a biased manner. One compelling idea

771   is that the lack of negative teaching signals in DLS plays a role in skill/habit, although further

772   examination is needed to establish its functions. Although we designed the task to minimize

773   effects of movement itself on results, accumulating studies suggested close relationship between

774   dopamine signaling and movement (Howe and Dombeck, 2016; da Silva et al., 2018). It is

775   important to test these other parameters in the future in order to understand the meaning of the

776   diversity of dopamine neurons and organization of dopamine-striatum systems.

777

778

779

780

781

782 **EXPERIMENTAL PROCEDURES**

783

784 **Animals**

785 17 dopamine transporter (DAT)-cre (B6.SJL-Slc6a3tm1.1(cre)Bkmn/J, Jackson Laboratory;

786 RRID:IMSR JAX:006660) (Bäckman et al., 2006) heterozygous mice, and 5 DAT-Cre;Ai14

787 (Rosa-CAG-LSL-tdTomato, Jackson Laboratory; RRID:IMSR JAX:007914) (Madisen et al.,

788 2010) double heterozygous mice, male and female, were used for recording signals from

789 dopamine axons. All mice were backcrossed with C57BL/6J (Jackson Laboratory). Animals

790 were housed on a 12 hour dark/12 hour light cycle (dark from 07:00 to 19:00) and performed a

791 task at the same time each day. All procedures were performed in accordance with the National

792 Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the

793 Harvard Animal Care and Use Committee.

794

795 **Surgical Procedures**

796 All surgeries were performed under aseptic conditions with animals anesthetized with isoflurane

797 (1–2% at 0.5–1.0 l/min). Analgesia was administered pre (buprenorphine, 0.1 mg/kg, I.P) and

798 postoperatively (ketoprofen, 5 mg/kg, I.P). To express GCaMP7f (Dana et al., 2019) specifically

799 in dopamine neurons, we unilaterally injected 300 nl of mixed virus solution; AAV5-CAG-

800 FLEX-GCaMP7f (1×10^{12} particles/ml, UNC Vector Core, NC) and AAV5-CAG-FLEX-

801 tdTomato (2×10^{13} particles/ml, UNC Vector Core, NC) into both the VTA and SNc (600 nl

802 total) in the DAT-cre mice. Only AAV5-CAG-FLEX-GCaMP7f (300 nl total) was used for

803 DAT;Ai14 double transgenic mice. Virus injection lasted around 20 minutes, and then the

804 injection pipette was slowly removed over the course of several minutes to prevent damage to

805 the tissue. We also implanted optic fibers (400 μm diameter, Doric Lenses, Canada) into the VS,

806 DMS, or DLS (1 fiber per mouse). To do this, we first slowly lowered optical fibers into the

807 striatum. Once fibers were lowered, we first attached them to the skull with UV-curing epoxy

808 (NOA81, Thorlabs, NJ), and then a layer of rapid-curing epoxy to attach the fiber cannulas even

809 more firmly to the underlying glue. After waiting 15 minutes for this to dry, we applied a black

810 dental adhesive (Ortho-Jet, Lang Dental, IL). We used magnetic fiber cannulas (Doric Lesnses,

811 MFC_400/430) and the corresponding patch cords to allow for recordings in freely moving

27

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,

28

843   mixtures of odor A and B (100/0, 90/10, 65/35, 35/65, 10/90, 0/100) were gradually introduced.

844   After the accuracy of all the mixture odors achieved more than 50%, neuronal recording with

845   fiber fluorometry was performed for 5 sessions. Subsequently, a task with different amounts of

846   water was introduced. Mixtures of odor A and B (100/0, 65/35, 35/65, 0/100) but no odor C were

847   used in this task. Each recording session started with 88-120 trials with an equal amount of water

848   (~6 ul, the standard amount) in the first block to calibrate any potential bias on the day. In the

849   second block, different amounts of reward were delivered in each water port. In order to make

850   the water amounts unpredictable, one water port delivered big or medium size of water (2.2 and

851   0.8 times of the standard, ~13.2 and 4.8 µl, BIG side) in a pseudo-random order, and another

852   water port delivered medium or small size of water (0.8 and 0.2 times of the standard, ~4.8 and

853   1.2 µl, SMALL side) in a pseudo-random order. Block 2 continued for 200 trials or until the end

854   of recording sessions, whichever came earlier. A mouse performed 134.3 ± 3.4 (mean $\pm$ SEM)

855   trials in block 2. The water condition (BIG or SMALL) was assigned to a left or right water port

856   in a pseudo-random order across sessions. Recording was conducted for 40 min every other day

857   to avoid potential bleaching. On days with no recording, animals were trained with pure odors A

858   and B with the standard amount of water.

859

860   **Fiber photometry**

861   Fiber fluorometry (photometry) was performed as previously reported (Menegas et al., 2018)

862   with a few modification. The optic fiber (400 µm diameter, Doric Lenses) allows chronic, stable,

863   minimally disruptive access to deep brain regions and interfaces with a flexible patch cord (Doric

864   Lenses, Canada) on the skull surface to simultaneously deliver excitation light (473 nm,

865   Laserglow Technologies, Canada; 561 nm, Opto Engine LLC, UT) and collect GCaMP and

866   tdTomato fluorescence emissions. Activity-dependent fluorescence emitted by cells in the

867   vicinity of the implanted fiber's tip was spectrally separated from the excitation light using a

868   dichroic, passed through a single band filter, and focused onto a photodetector connected to a

869   current preamplifier (SR570, Stanford Research Systems, CA). During recording, optic fibers

870   were connected to a magnetic patch cable (Doric Lesnses, MFP_400/430) which delivered

871   excitation light (473 nm and 561 nm) and collected all emitted light. The emitted light was

872   subsequently filtered using a 493/574 nm beam-splitter (Semrock, NY) followed by a 500 ± 20

873   nm (Chroma, VT) and 661 ± 20 nm (Semrock, NY) bandpass filters and collected by a

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, $50-b_0/b_1$ 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

905 100s. Then, the correlation between green and red signals during ITI was examined by linear

906 regression. If the correlation is significant (p<0.05), fitted tdTom signals were subtracted from

907 green signals.

908

909 Responses were calculated by subtracting the average baseline activity from the average activity

910 of the target window. Unless specified otherwise, odor responses were calculated by averaging

911 activity from 1-0 s before odor port out (before choice) minus the average activity from the

912 baseline period (1-0.2 s before odor onset). Responses after choice were calculated by averaging

913 activity from 0-1 s after water port in minus the same baseline. Outcome responses were

914 calculated by averaging activity from 0-1 s after water onset minus the same baseline. When

915 comparing activity before and after water onset, average activity in 1-0.2 s before water onset

916 was used as baseline. To normalize GCaMP signals across sessions within an animal, GCaMP

917 signals were divided by average of peak responses during 1 s after odor onset in all the

918 successful trials in the session. Z-scores of the signals were obtained using mean and standard

919 deviation of signals in all the choice trials (from 2 s before odor onset to 6 s after odor onset) in

920 each animal.

921

922 We built a regularized linear regression to fit cosine kernels (Park et al., 2014) (width of 200 ms,

923 interval of 40 ms) to the activity of dopamine axons in each animal. We used down-sampled

924 (every 20 ms) responses in all valid choice trials (trials with >1s odor sampling time and any

925 choice, -1 to 7 s from odor onset) for the model fitting. We used 4 different time points to lock

926 kernels: odor onset ("odor"), odor port out ("movement"), water port in ("choice"), and water

927 onset ("water"). Odor kernels consist of 4 types of kernels: "base" kernels to span -960 to 200 ms

928 from odor onset in all trials, and "pure big" kernels in trials with a pure odor associated with

929 big/medium water, "pure small" kernels in trials with a pure odor associated with medium/small

930 water, and "mixture" kernels in trials with a mixture odor to span 0-1600 ms from odor onset.

931 Movement kernels consist of 2 types of kernels: "contra turn" kernels in trials with choice contra-

932 lateral to the recording site, and "ipsi turn" kernels in trials with choice ipsi-lateral to the

933 recording site to span -1000 to 1200 ms from when a mouse exited an odor port. Choice kernels

934 consist of 3 types of kernels: "correct big" kernels in trials with correct choice of medium/small

935 water and "correct small" kernels in trials with correct choice of medium/small water to span -

31

936   400 to 1200 ms from when a mouse entered a water port (water port in), and "error" kernels in

937   trials with choice error to span -400 to 5200 ms from water port in. Water kernels consist of 4

938   types of kernels: "big water" kernels for big size of water, "medium water big side" kernels for

939   medium size of water at a water port of big/medium water, "medium water small side" kernels

940   for medium size of water at a water port of medium/small water, and "small water" for small size

941   of water to span 0-4200 ms after water onset. All the kernels were fitted to responses using linear

942   regression with Lasso regularization with 10-fold cross validation. Regularization coefficient

943   lambda was chosen so that cross-validation error is minimum plus one standard deviation. %

944   explained by a model was expressed as reduction of a variance in the residual responses

945   compared to the original responses. Contribution of each component in the model was measured

946   by reduction of a deviance compared to a reduced model excluding the component.

947

948   We estimated response function to water in dopamine axons with linear regression with power

949   function in each animal.

950
$$r \; = \; k(R^{\alpha} \; + \; c1 \; * \; S \; + \; c2)$$

951   where r is the dopamine axon response to water, R is the water amount, S is SMALL side (S=1

952   when water was delivered at SMALL side, S=0 otherwise). There are 4 different conditions,

953   responses to big and medium water at a port of BIG side, and to medium and small water at a

954   port of SMALL side. We first optimized α by minimizing average of residual sum of squares for

955   each animal and then applied α =0.7 for all the animals to obtain other parameters, k, c1, and c2.

956   The response function was drawn with R as x-axis and r as y-axis. The amount of water to which

957   dopamine axons do not respond under expectation of BIG or SMALL water was estimated by

958   getting a crossing point of the obtained response function where the value is 0 (a zero-crossing

959   point). The distribution of zero-crossing points was examined by linear regression of zero-

960   crossing values against anatomical locations (anterior-posterior, dorsal-ventral, and medial-

961   lateral). To visualize zero crossing points on the atlas, zero-crossing values were fitted against

962   anatomical locations with interaction terms using linear regression with elastic net regularization

963   (α=0.1) with 3-fold cross validation. The constructed map was sliced at a coronal plane Bregma

964   +0.7 and overlaid on an atlas (Paxinos and Franklin, 2019).

965

966   To visualize activity pattern in multiple time windows at the same time, we stretched activity in

967   each trial to standard windows. Standard windows from odor onset to odor poke out, and from

968   odor poke out to water poke in, were determined by median reaction time and median movement

969   time for each animal. For average plots of multiple animals, windows were determined by the

970   average of median reaction times and of median movement times in all animals. The number of

971   100ms bins in each time window was determined by dividing median reaction time and median

972   movement time by 100. Dopamine responses in the window were divided into the bin number

973   and the average response in each bin was stretched to 100ms. The stretched activity patterns

974   were used only for visualization, and all the statistical analyses were performed using original

975   responses.

976

977   **Estimation of state values and TD errors using simulations**

978   To examine how the value and RPE may change within a trial, we employed a Monte-Carlo

979   approach to simulate animal's choices at a steady state (i.e. after the animal learned the task). We

980   used a Monte-Caro approach to obtain the *ground truth* state values as the animal progresses

981   through task events without assuming a specific learning algorithm, under the assumption that

982   the animal has learned the task. After obtaining the state values, we computed TD errors over the

983   obtained state values.

984

985   *Model architecture*

986   We considered two types of models. The variability and errors in choice in psychophysical

987   performance can arise from at least two noise sources; noise in the variability in the process of

988   estimating sensory inputs (perceptual noise) and noise in the process of selecting an action

989   (decision noise). The first model contained only perceptual noise (Green and Swets, 1966), and

990   the second model contained both perceptual and decision noise.

991

992   These models had different "states" considering N_S subjective odors ($N_S = 60$ or 4 discrete

993   states), choice (BIG versus SMALL), and different timing (inter-trial interval, odor port entry,

994   odor presentation, choice, water port in, waiting for reward, and receiving feedback/outcome)

995   (circles in Figure 7A).

996

997 We assumed N_S possible subjective odor states (O') which comprise SubOdor1 and SubOdor2

998 states. We assumed that, in each trial, an internal estimate of the stimulus or a "subjective odor"

999 (O') was obtained by adding a noise to the presented odor stimulus (O) (one of the 4 mixtures of

1000 Odor A and B; 100/0, 65/35, 35/65, 0/100) (Figure 7A-C). In the model, the probability of falling

1001 on a given subjective odor state (O') is calculated using a Gaussian distribution centering on the

1002 presented odor (O) with the standard deviation, σ. We considered two successive states for

1003 subjective odor states in order to reflect a relatively long duration before an odor port exit.

1004

1005 As in the behavioral paradigm, whether the model receives a reward or not was determined

1006 solely by whether the presented odor (O) instructed the BIG side or SMALL side. Each

1007 subjective odor state contains cases when the presented odor (O) is consistent or congruent with

1008 the subjective odor (O'). For each subjective odor state, the probability of receiving a reward

1009 after choosing the BIG side, $p(BIG\ is\ correct) = f_B$, can be calculated as the fraction of cases

1010 when the presented odors instructed the BIG side. Conversely, the probability of reward after

1011 choosing the SMALL side is $p(SMALL\ is\ correct) = f_S = 1 - f_B$. Note that neither f_B nor f_S

1012 depends on reward size manipulations (as will be discussed later, the animal's choices will be

1013 dependent on reward size manipulations).

1014

1015 *Action selection*

1016 For each subjective odor, the model chose either the BIG or the SMALL side based on the value

1017 of choosing the BIG or SMALL side (V_B and V_S respectively, equivalent to the state value of the

1018 next state after committing to choose the BIG or SMALL side; see below for how V_B and V_S

1019 were obtained). In the first model which contains only perceptual noise, the side that is

1020 associated with a larger value is chosen. In the second model which contains both perceptual and

1021 decision noise, a choice is made by transforming V_B and V_S into the probability of choosing a

1022 given option using a sigmoidal function (e.g. Boltzmann distribution) (Sutton and Barto, 2011).

1023 In the softmax, the probabilities of choosing the BIG and SMALL side (P_B, P_S) are given,

1024 respectively, by,

1025
$$P_B = \frac{e^{(V_B/(V_B+V_S))/\tau}}{e^{(V_B/(V_B+V_S))/\tau} + e^{(V_S/(V_B+V_S))/\tau}}$$

1026
$$P_S = 1 - P_B$$

1027 We also tested other stochastic decision rules such as Herrnstein's matching law (Herrnstein,

1028 1961) or ε-greedy exploration (randomly selecting an action in a certain fraction [ε] of trials)

1029 (Sutton and Barto, 2011). In Herrnstein's matching law, the probability of choosing the BIG side

1030 is given by,

1031
$$P_B = \frac{V_S}{V_S + V_B}$$

1032

1033 The perceptual noise and a set of decision rule determine the behavioral performance of the

1034 model. The first model has only one free parameter, σ. The second model has one or no

1035 additional parameter (τ for softmax, or ε, for ε-greedy; no additional parameter for matching).

1036 We first obtained the best fit parameter(s) based on the behavioral performance of all animals

1037 (the average performance in Block 2; i.e. Figure 1C, orange) by minimizing the mean squared

1038 errors in the psychometric curves.

1039

1040 For the first model, the best fit σ was 21% Odor. We also tested with σ of 5%, and the TD error

1041 dynamic was qualitatively similar. For the second model using the softmax rule, the best fit τ

1042 was 0.22 while σ was 18% Odor.

1043

1044 *State values*

1045 The state value for each state was obtained as the weighted sum of expected values of available

1046 options which was computed by multiplying expected values of the option with probability of an

1047 option in the next step.

1048

1049 Outcome2 state represents the timing when the animal recognizes the amount of water. The state

1050 value is given by the amount of water that the model received (big, medium, small),

1051
$$V_b = 2.2^\alpha$$

1052
$$V_m = 0.8^\alpha$$

1053
$$V_s = 0.2^\alpha$$

1054 where the exponent $\alpha = 0.7$ makes the value function a concave function of reward amounts,

1055 similar to the fitting analysis of the fluorometry data (Figure 4C). Using $\alpha = 1$ (i.e. a linear

1056 function) did not change the results.

35

1057

1058   Ourcome1 state, or Water/No-water states (W and N, respectively) represent when the animal

1059   noticed the presence or absence of reward, respectively, but not the amount of reward. The value

1060   of a W (Water) state was defined by the average value of the next states. At the BIG side,

1061 $$V_{WB} = (V_b + V_m)/2$$

1062

1063   whereas at the SMALL side,

1064 $$V_{WS} = (V_m + V_s)/2$$

1065   The values of N (No-water) states at the BIG and SMALL side are zero,

1066 $$V_{NB} = 0$$

1067 $$V_{NS} = 0$$

1068

1069   WaterPort1 and WaterPort2 states represent when the animal entered and stayed in the water

1070   port, respectively. The state value was obtained separately for the BIG and SMALL side. The

1071   value of choosing the BIG and SMALL sides is given by weighted sum of the values of the next

1072   states ($V_{WB}, V_{NB}, V_{WS}, V_{NS}$). The probabilities of transiting to the W and N states are given by the

1073   probability of receiving a reward given the choice (BIG or SMALL). As discussed above, these

1074   probabilities are given by f_B and f_S, respectively. Thus,

1075 $$V_B = f_B \cdot V_{WB}$$

1076 $$V_S = f_S \cdot V_{WS}$$

1077   We considered two successive states for WaterPort states to reflect a relatively long duration

1078   before receiving feedback/outcome. The two successive states had the same state values.

1079

1080   SubOdor1 and SubOdor2 states represent when the animal obtained a subjective odor (O') and

1081   before making a choice. The model chooses the BIG or SMALL side with the probability of

1082 P_B and P_S , respectively, as defined above. Therefore, the state value of WaterPort1 and

1083   WaterPort2 was defined by the weighted sum of the values of the next states (V_B and V_S),

1084 $$V_{O'} = P_B V_B + P_S V_S$$

1085   The two successive states had the same state values.

1086

1087   OdorOn state represents when the animal recognized the presentation of an odor but before

1088   recognizing the identity of that odor. The state value of the OdorOn state is defined by the

1089   weighted sum of the values of the next states (SubOdor1).

1090

1091   ITI state represents when the animal is in the inter-trial interval (i.e. before odor presentation).

1092   The value of ITI state was set to zero.

1093

1094   *TD errors*

1095   After obtaining state values at each state, we then computed TD errors using a standard

1096   definition of TD error which is the difference between the state values at consecutive time points

1097   plus received rewards at each time step (Sutton and Barto, 1987). For simplicity, a discounting

1098   factor was set to 1 (no discounting).

1099

1100   *Invalid trials*

1101   We also tested the effect of including invalid trials. At water acquisition, we included failures

1102   (20% of trials, value 0) where a mouse did not fulfil the requirement of odor poke duration (short

1103   odor poke), but did indicate a choice. At an odor port, failures resulted from multiple pokes of

1104   odor port (4% of trials), and a short odor poke (14% of trials). Values for these failures were set

1105   to 0. Existence or omission of these failures in models did not change the conclusion.

1106

1107   **Randomization, blinding, and data exclusion**

1108   Chemicals were randomly assigned to an odor cue. Trial types (odors) were pseudorandomized

1109   in a block. Session types were pseudorandomized in a recording schedule. Animals were

1110   randomly assigned to a recording location. The experimenter did not know location of recording

1111   until the recording schedule was completed. No animals were excluded from the study: all

1112   analysis includes data from all animals. No trials were excluded from statistical analyses. To

1113   visualize average activity pattern in a stretched time-window, outlier trials (maximum, minimum

1114   or average activity of a trial is outside of $3 \times$ standard deviation of maximum, minimum or

1115   average activity of all the trials) were excluded.

1116

1117   **Statistical analyses**

1118 Data analysis was performed using custom software written in Matlab (MathWorks, Natick, MA,

1119 USA). All code used for analysis is available on request. All statistical tests were two-sided. For

1120 statistical comparisons of the mean, we used one-way ANOVA and two-sample Student's t tests,

1121 unless otherwise noted. Paired t tests were conducted when the same mouse's neural activity was

1122 being compared across different conditions or different time windows. The significance level

1123 was corrected for multiple comparisons using Holm–Sidak's tests unless otherwise indicated. All

1124 error bars in the figures are s.e.m. In boxplots, the edges of the boxes are the 25th and 75th

1125 percentiles (q1 and q3, respectively), and the whiskers extend to the most extreme data points not

1126 considered outliers. Points are drawn as outliers if they are larger than $q3+1.5\times(q3-q1)$ or $q1-$

1127 $1.5\times(q3-q1)$. Individual data points were overlaid on boxplots to compare striatal areas.

1128

Author Contributions:

1129
1130
1131   ITK and MWU designed experiments and collected and analyzed data. HM performed pilot experiments.
1132   The results were discussed and interpreted by ITK, NU and MWU. ITK, NU and MWU wrote the paper.
1133
1134
1135

Acknowledgments

1145
1146

Declaration of Interests

1148
1149   The authors declare no competing interests.
1150
1151
1152

39

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.

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.

41

1217 Herrnstein, R.J. (1961). Relative and absolute strength of responses as a function of frequency
1218 of reinforcement.

1219 Hikosaka, O., Rand, M.K., Miyachi, S., and Miyashita, K. (1995). Learning of sequential
1220 movements in the monkey: process of learning and retention of memory. J. Neurophysiol. *74*,
1221 1652–1661.

1222 Hirokawa, J., Vaughan, A., Masset, P., Ott, T., and Kepecs, A. (2019). Frontal cortex neuron types
1223 categorically encode single decision variables. Nature *576*, 446–451.

1224 Hong, S., Jhou, T.C., Smith, M., Saleem, K.S., and Hikosaka, O. (2011). Negative reward signals
1225 from the lateral habenula to dopamine neurons are mediated by rostromedial tegmental
1226 nucleus in primates. J. Neurosci. *31*, 11457–11471.

1227 Howe, M.W., and Dombeck, D.A. (2016). Rapid signalling in distinct dopaminergic axons during
1228 locomotion and reward. Nature *535*, 505–510.

1229 Hunnicutt, B.J., Jongbloets, B.C., Birdsong, W.T., Gertz, K.J., Zhong, H., and Mao, T. (2016). A
1230 comprehensive excitatory input map of the striatum reveals novel functional organization. ELife
1231 *5*.

1232 Ilango, A., Kesner, A.J., Keller, K.L., Stuber, G.D., Bonci, A., and Ikemoto, S. (2014). Similar Roles
1233 of Substantia Nigra and Ventral Tegmental Dopamine Neurons in Reward and Aversion. J.
1234 Neurosci. *34*, 817–822.

1235 Jhou, T.C., Geisler, S., Marinelli, M., Degarmo, B.A., and Zahm, D.S. (2009a). The mesopontine
1236 rostromedial tegmental nucleus: a structure targeted by the lateral habenula that projects to
1237 the ventral tegmental area of Tsai and substantia nigra compacta. J. Comp. Neurol. *513*, 566–
1238 596.

1239 Jhou, T.C., Fields, H.L., Baxter, M.G., Saper, C.B., and Holland, P.C. (2009b). The rostromedial
1240 tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes
1241 aversive stimuli and inhibits motor responses. Neuron *61*, 786–800.

1242 de Jong, J.W., Afjei, S.A., Pollak Dorocic, I., Peck, J.R., Liu, C., Kim, C.K., Tian, L., Deisseroth, K.,
1243 and Lammel, S. (2019). A Neural Circuit Mechanism for Encoding Aversive Stimuli in the
1244 Mesolimbic Dopamine System. Neuron *101*, 133-151.e7.

1245 Kamin, L.J. (1969). Predictability, surprise, attention and conditioning. Punishm. Aversive Behav.

1246 Keiflin, R., Pribut, H.J., Shah, N.B., and Janak, P.H. (2019). Ventral Tegmental Dopamine Neurons
1247 Participate in Reward Identity Predictions. Curr. Biol. *29*, 93-103.e3.

1248 Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation
1249 and behavioural impact of decision confidence. Nature *455*, 227–231.

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.

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).

1315   Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness
1316   of conditioned but not of unconditioned stimuli. Psychol. Rev. *87*, 532.

1317   Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology
1318   of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

1319   Rausch, M., and Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical
1320   signature of decision confidence. PLOS Comput. Biol. *15*, e1007456.

1321   Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the
1322   effectiveness of reinforcement and nonreinforcement. Class. Cond. II Curr. Res. Theory *2*, 64–
1323   99.

1324   Robbins, T.W., and Costa, R.M. (2017). Habits. Curr. Biol. *27*, R1200–R1206.

1325   Rorie, A.E., Gao, J., McClelland, J.L., and Newsome, W.T. (2010). Integration of Sensory and
1326   Reward Information during Perceptual Decision-Making in Lateral Intraparietal Cortex (LIP) of
1327   the Macaque Monkey. PLOS ONE *5*, e9308.

1328   Sajad, A., Godlove, D.C., and Schall, J.D. (2019). Cortical microcircuitry of performance
1329   monitoring. Nat. Neurosci. *22*, 265–274.

1330   Samejima, K., and Doya, K. (2007). Multiple Representations of Belief States and Action Values
1331   in Corticobasal Ganglia Loops. Ann. N. Y. Acad. Sci. *1104*, 213–228.

1332   Saunders, B.T., Richard, J.M., Margolis, E.B., and Janak, P.H. (2018). Dopamine neurons create
1333   Pavlovian conditioned stimuli with circuit-defined motivational properties. Nat. Neurosci. *21*,
1334   1072–1083.

1335   Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward.
1336   Science *275*, 1593–1599.

1337   da Silva, J.A., Tecuapetla, F., Paixão, V., and Costa, R.M. (2018). Dopamine neuron activity
1338   before action initiation gates and invigorates future movements. Nature *554*, 244–248.

1339   Smith, K.S., and Graybiel, A.M. (2016). Habit formation. Dialogues Clin. Neurosci. *18*, 33–43.

1340   Stuphorn, V., Taylor, T.L., and Schall, J.D. (2000). Performance monitoring by the supplementary
1341   eye field. Nature *408*, 857–860.

1342   Suri, R.E., and Schultz, W. (1999). A neural network model with dopamine-like reinforcement
1343   signal that learns a spatial delayed response task. Neuroscience *91*, 871–890.

1344   Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. Mach. Learn. *3*,
1345   9–44.

1346   Sutton, R.S., and Barto, A.G. (1987). A temporal-difference model of classical conditioning. In
1347   Proceedings of the Ninth Annual Conference of the Cognitive Science Society, (Seattle, WA), pp.
1348   355–378.

1349   Sutton, R.S., and Barto, A.G. (2011). Reinforcement learning: An introduction.

1350   Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning, second edition: An Introduction
1351   (MIT Press).

1352   Thorndike, E.L. (1932). The fundamentals of learning (New York, NY, US: Teachers College
1353   Bureau of Publications).

1354   Tian, J., Huang, R., Cohen, J.Y., Osakada, F., Kobak, D., Machens, C.K., Callaway, E.M., Uchida, N.,
1355   and Watabe-Uchida, M. (2016). Distributed and Mixed Information in Monosynaptic Inputs to
1356   Dopamine Neurons. Neuron *91*, 1374–1389.

1357   Uchida, N., and Mainen, Z.F. (2003). Speed and accuracy of olfactory discrimination in the rat.
1358   Nat. Neurosci. *6*, 1224–1229.

1359   Wang, A.Y., Miura, K., and Uchida, N. (2013). The dorsomedial striatum encodes net expected
1360   return, critical for energizing performance vigor. Nat. Neurosci. *16*, 639–647.

1361   Watabe-Uchida, M., and Uchida, N. (2018). Multiple dopamine systems: Weal and woe of
1362   dopamine. In Cold Spring Harbor Symposia on Quantitative Biology, (Cold Spring Harbor
1363   Laboratory Press), pp. 83–95.

1364   Yetnikoff, L., Lavezzi, H.N., Reichard, R.A., and Zahm, D.S. (2014). An update on the connections
1365   of the ventral mesencephalic dopaminergic complex. Neuroscience *282*, 23–48.

1366   Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve
1367   outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. *19*,
1368   181–189.

1369   Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005). The role of the dorsomedial
1370   striatum in instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

1371

1372

1373

1374

1375

Figure 1
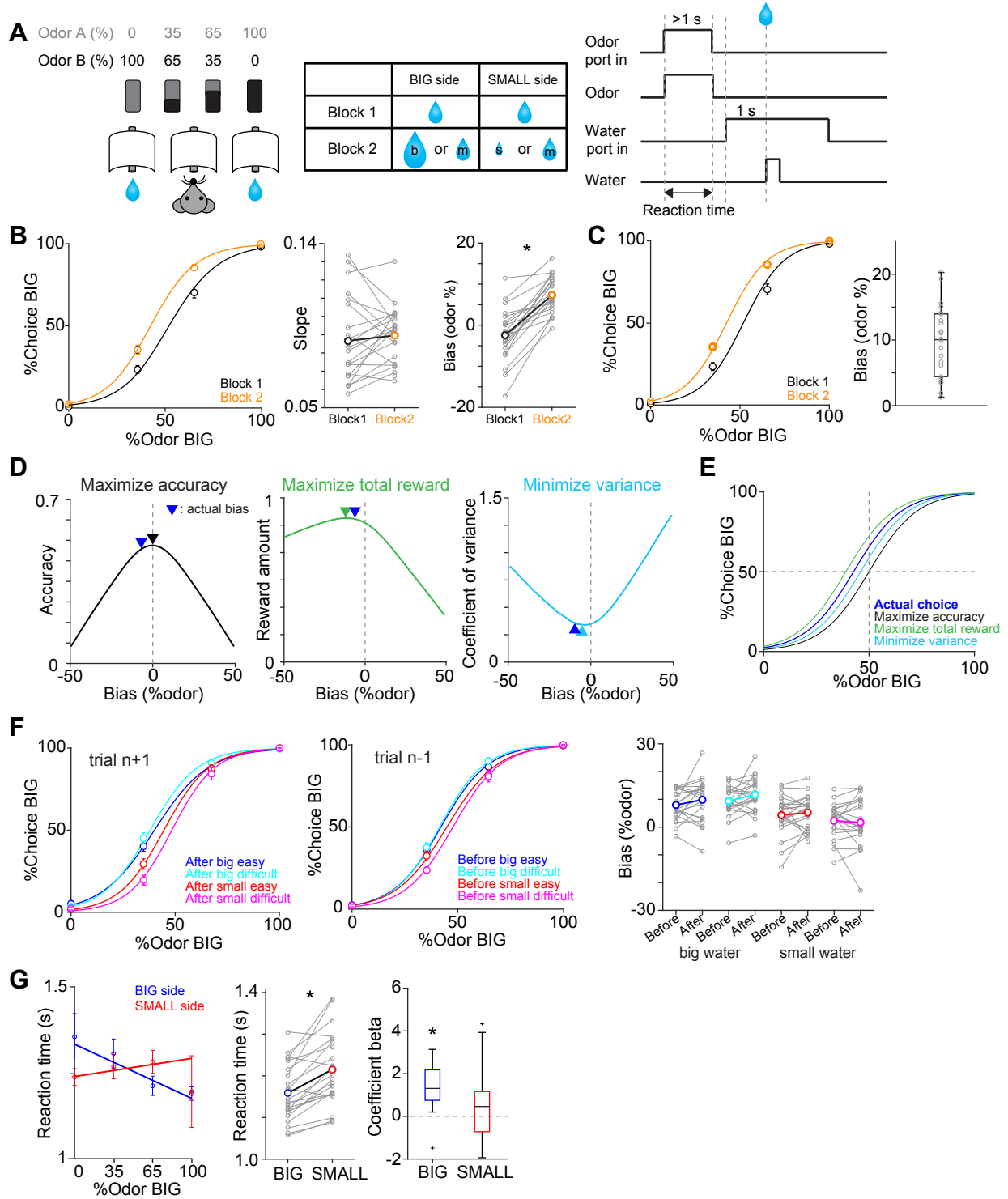
Figure 1. **Perceptual choice paradigm with probabilistic reward conditions** (A) A mouse discriminated a dominant odor in odor mixtures that indicates water availability in either the left or right water port. Correct choice was rewarded by a drop of water. In each session, an equal amount of water 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.

Figure 2
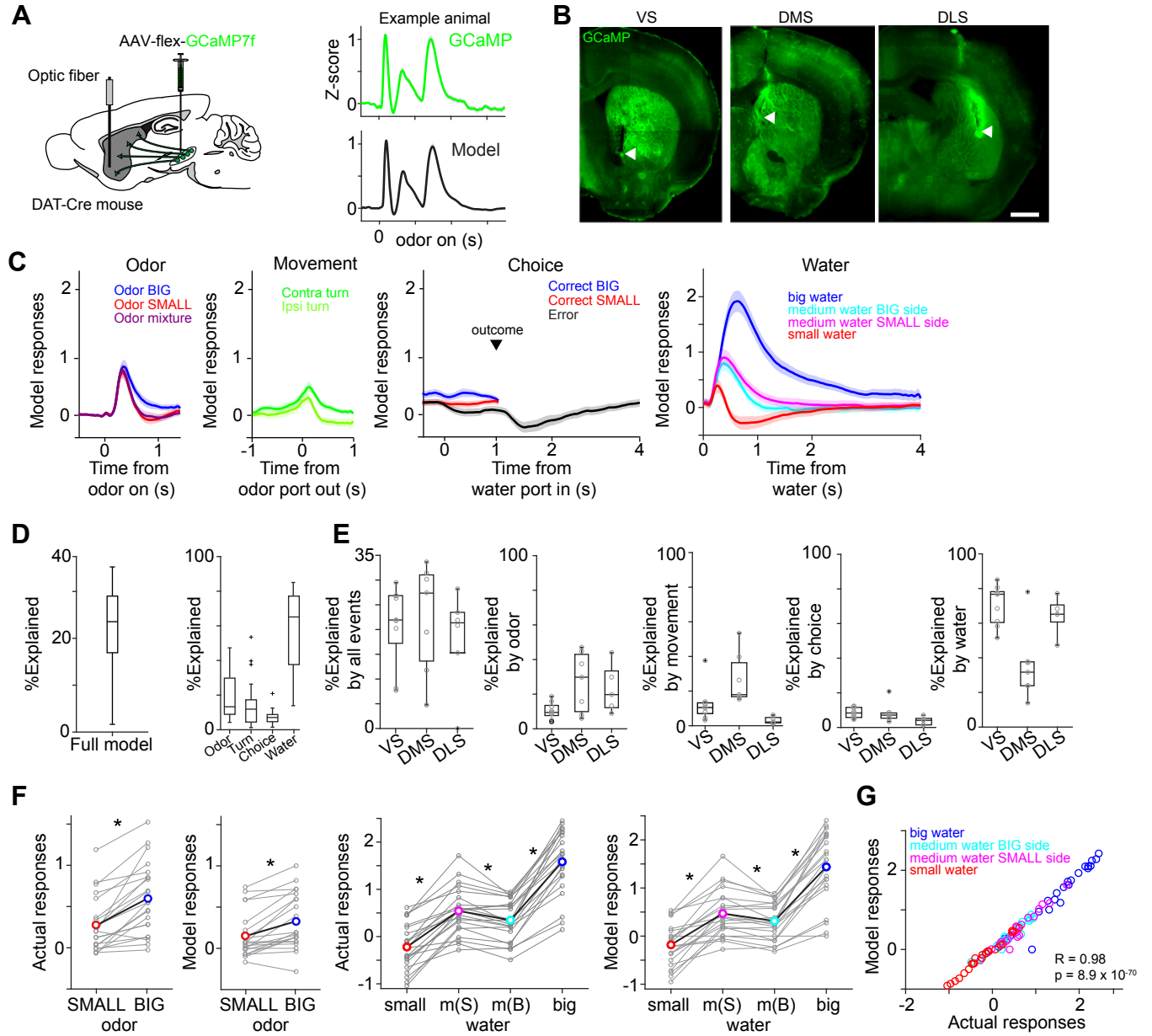
Figure 2. Dopamine axons in the striatum show characteristics of RPE (A) AAV-flex-GCaMP7f was injected in VTA and SNc, and dopamine axon activity was measured with an optic fiber inserted in the striatum. Right top, dopamine axon activity in all the valid trials (an animal chose an either water port 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.

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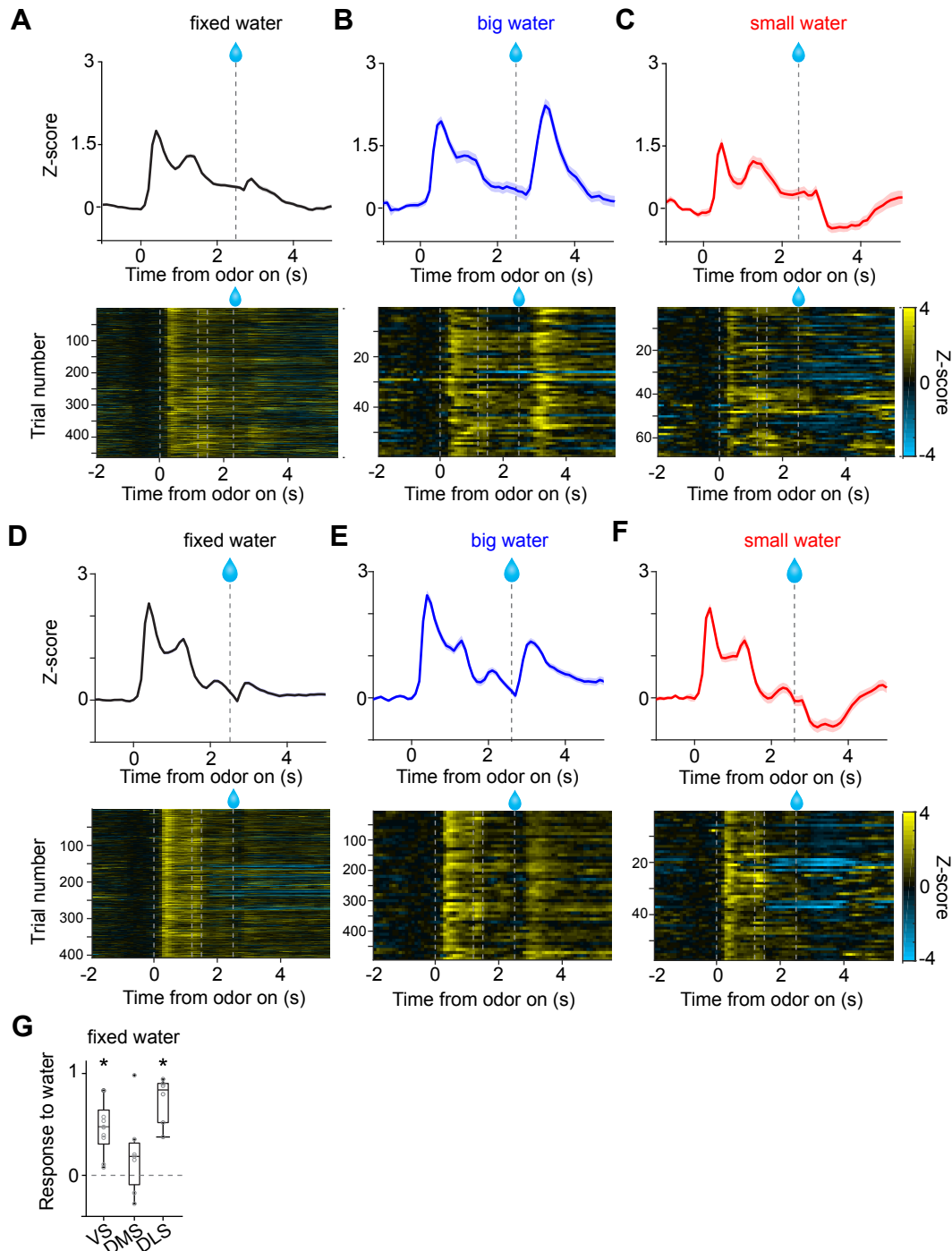
Figure 3. Small responses to fixed amounts of water in dopamine axons in DMS

(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=6.3×10^{-4}), but not in DMS (p=0.28).

Figure 4
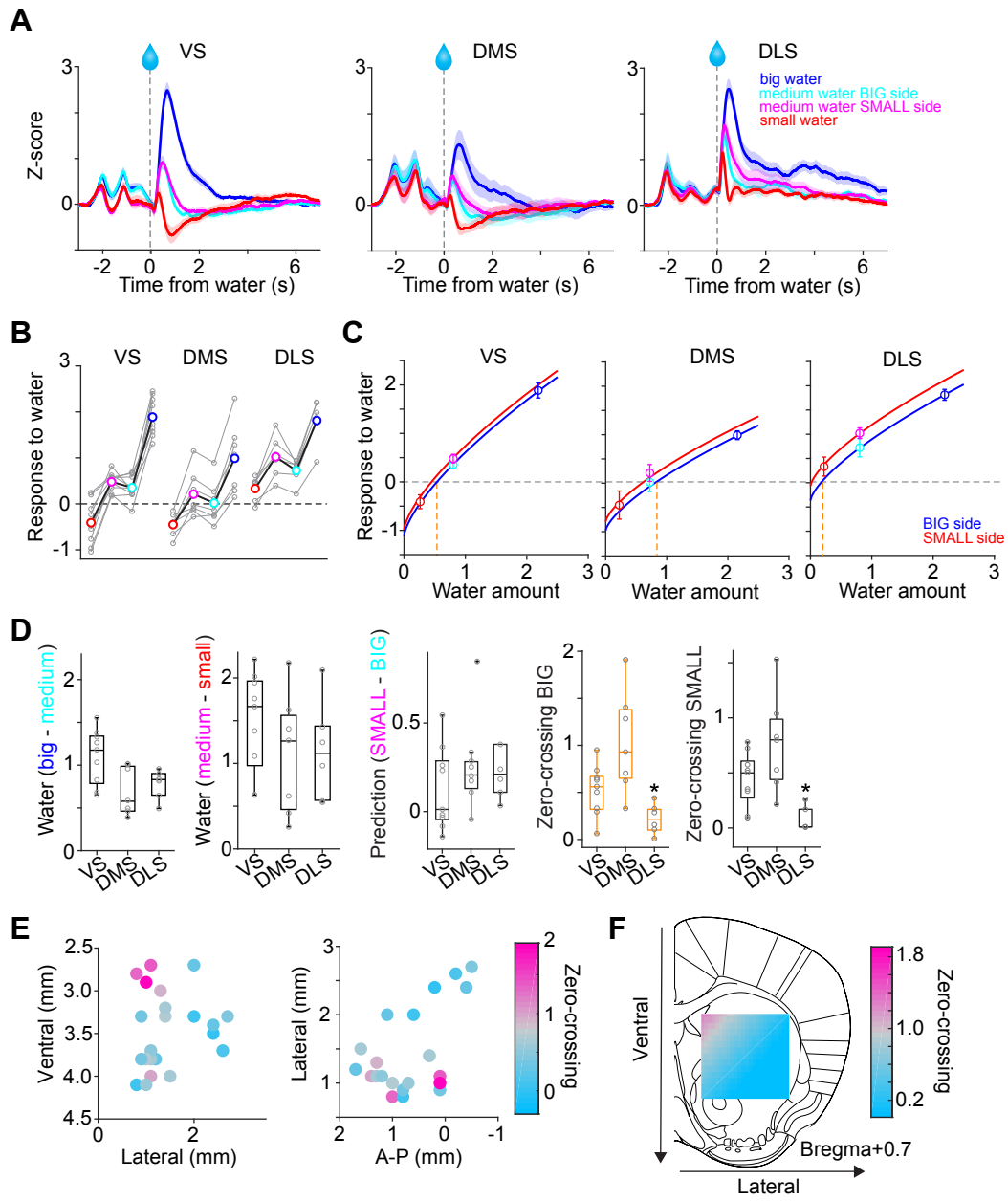
Figure 4. Responses to water in dopamine axons in the striatum (A) Activity patterns per different striatal location, aligned at water onset (mean ± SEM, n=9 for VS, n=7 for DMS, n=6 for DLS). (B) Average responses to each water condition in each animal grouped by striatal areas. (C) Average response 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).

Figure 5
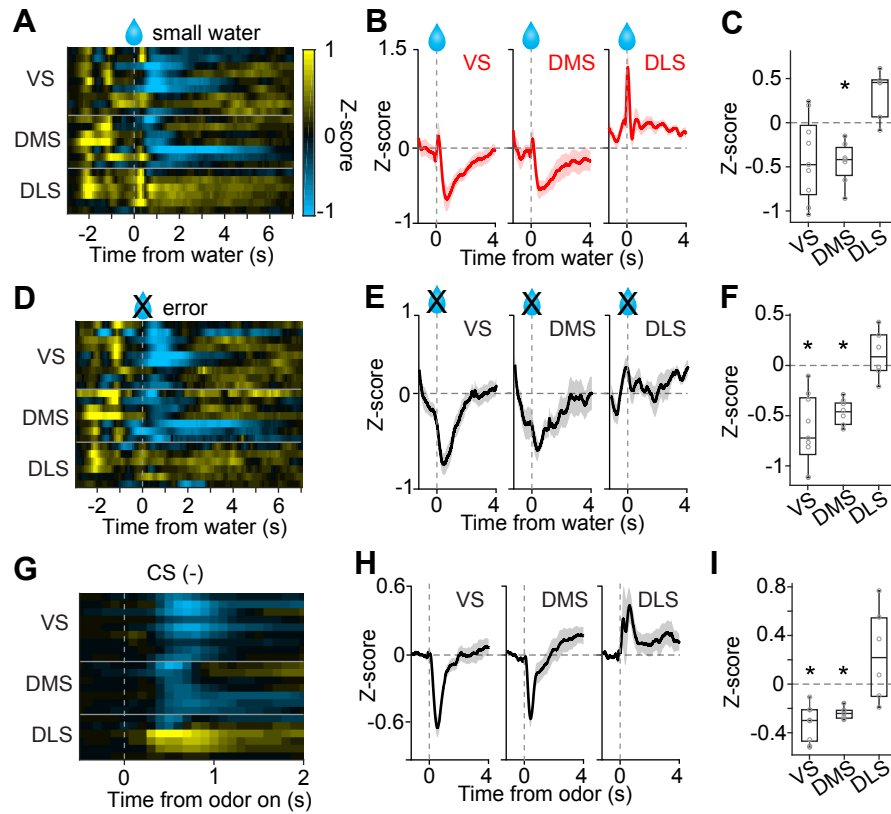
Figure 5. No inhibition by negative prediction error in dopamine axons in DLS (A) Activity pattern in each recording site aligned at small water. (B) Average activity pattern in each brain area (mean ± SEM). (C) Mean responses to small water (0.3-1.3 s after water onset) were negative in VS and DMS (p=0.031, 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=2.8×10^{-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=6.2×10^{-4}, p=6.5×10^{-5}, responses versus baseline), but not in DLS. The responses were different across striatal areas (p=1.5×10^{-4}, ANOVA; p=5.8×10^{-4}, VS versus DLS; p=1.6×10^{-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=1.4×10^{-4}, VS; p=1.0×10^{-5}, DMS, responses versus baseline), but not in DLS. Responses were different across striatal areas (p=2.5×10^{-4}, ANOVA; p=0.0012, VS versus DLS; p=0.0065, DMS versus DLS).

Figure 6
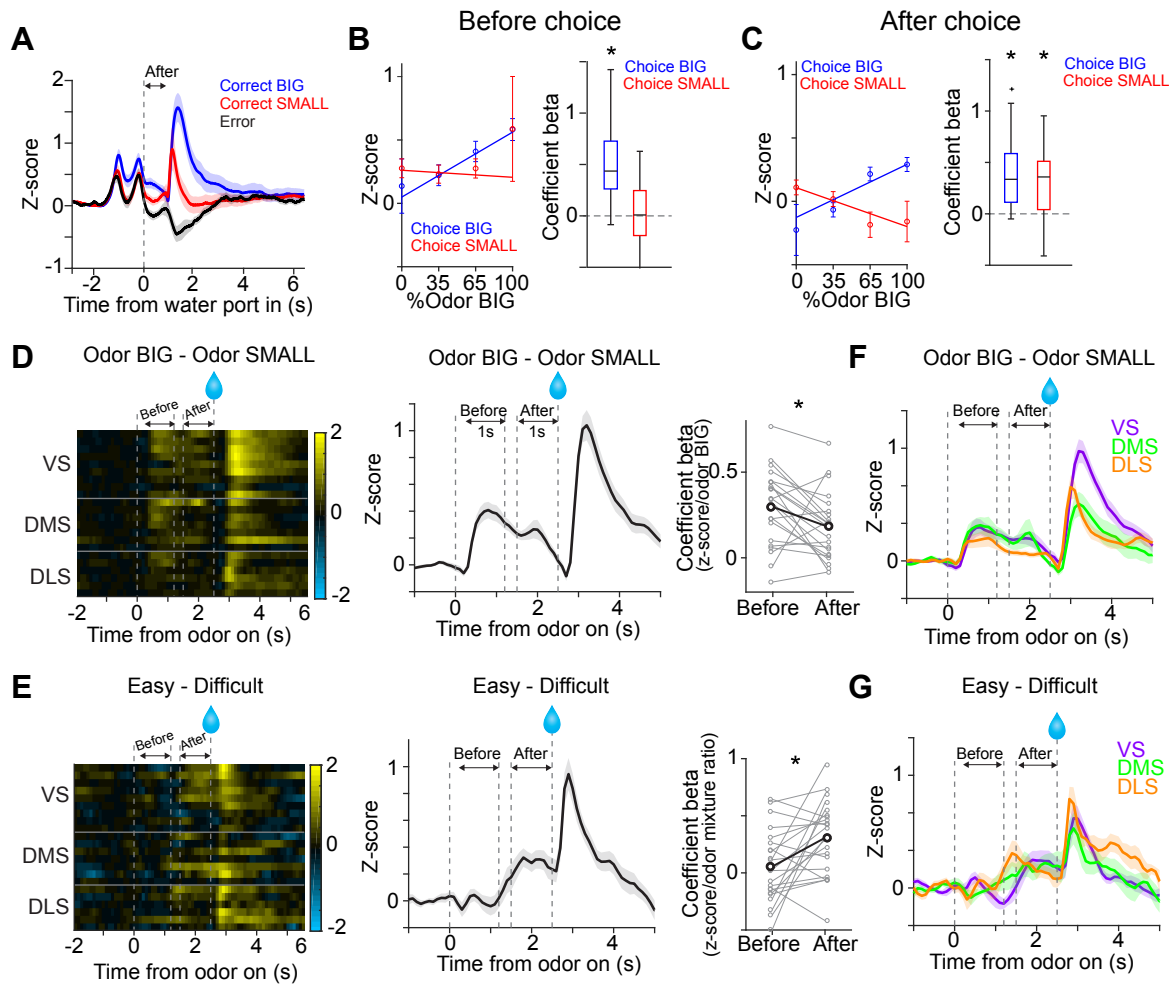
Figure 6. Dopamine signals stimulus-associated value and sensory evidence with different dynamics (A) Dopamine axon activity pattern aligned to time of water port entry for all animals (mean ± SEM). (B) Responses before choice (-1-0 s before odor port out) were fitted with linear regression with odor mixture ratio, and coefficient beta (slope) for all the animals are plotted. Correlation slopes were significantly positive for choice of the BIG side (p=5.6×10⁻⁶), 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=1.4×10⁻⁵) and of the SMALL side (p=2.2×10⁻⁴). (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 signi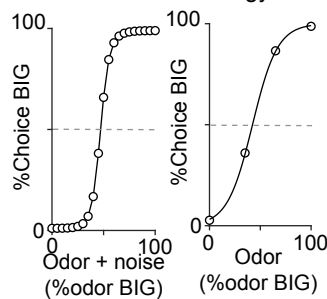ficantly different across areas (p=0.86, ANOVA). (G) Average difference in activity (easy minus difficult) in each striatal area. The difference of coefficient (before versus after choice) was not significantly different across areas (p=0.25, ANOVA).

Figure 7
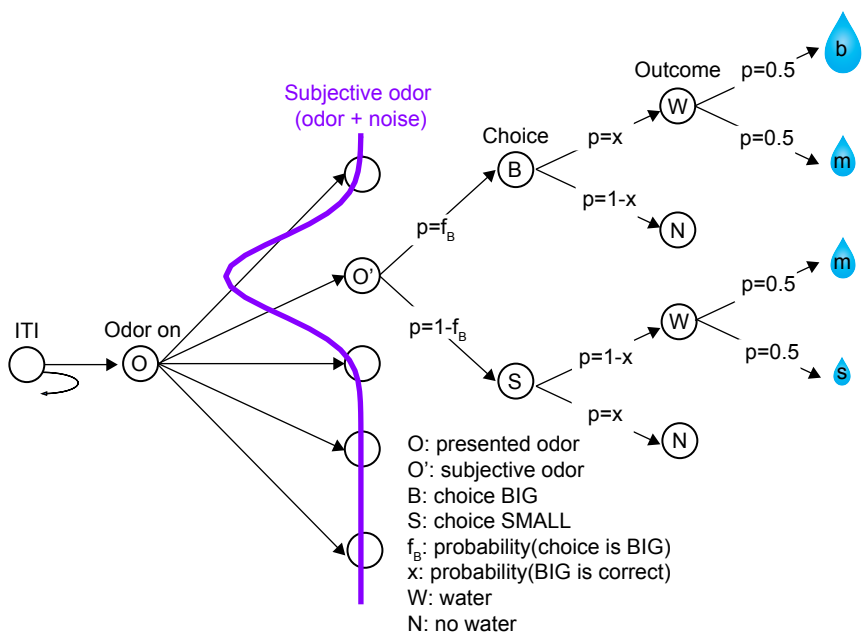
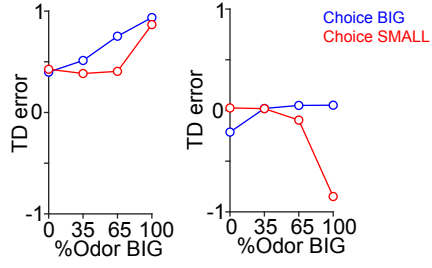**Figure 7. TD error dynamics capture emergence of sensory evidence after stimulus-associated value in dopamine axon activity** (A) Trial structure in the model. Some repeated states are omitted for clarification. (B-D) Models were constructed by adding perceptual noise with normal distribution to each 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 %).
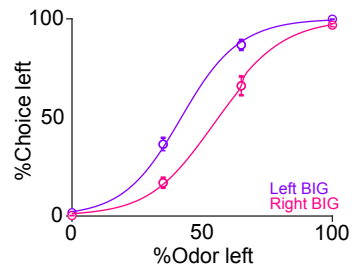
Figure S1

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.
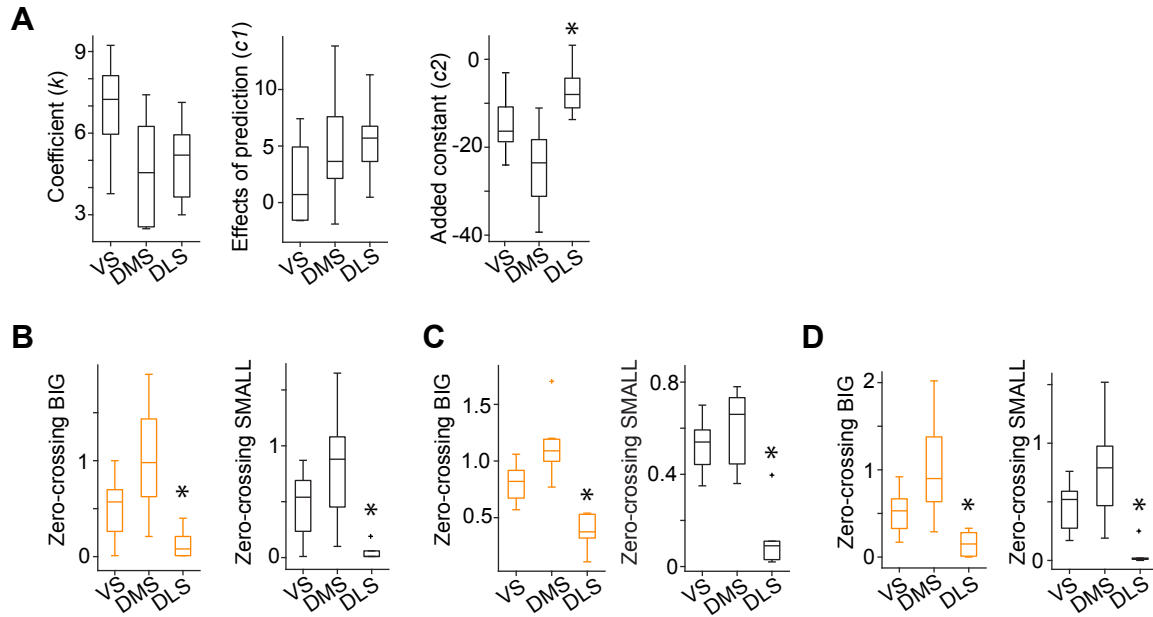
Figure S2

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=6.1x10^{-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=5.8x10^{-5} for BIG; p=2.1x10^{-4} for SMALL, ANOVA). (D) Zero-crossing points using kernel models with power function (p=0.0033 and p=8.9x10^{-4}, ANOVA).
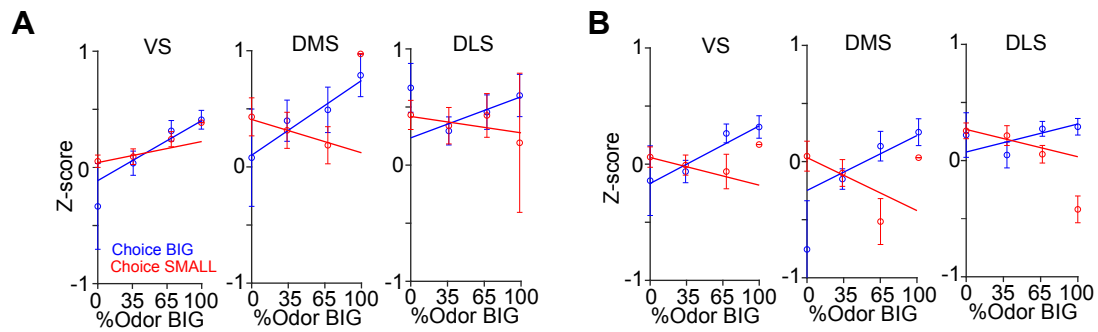
Figure S3

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).
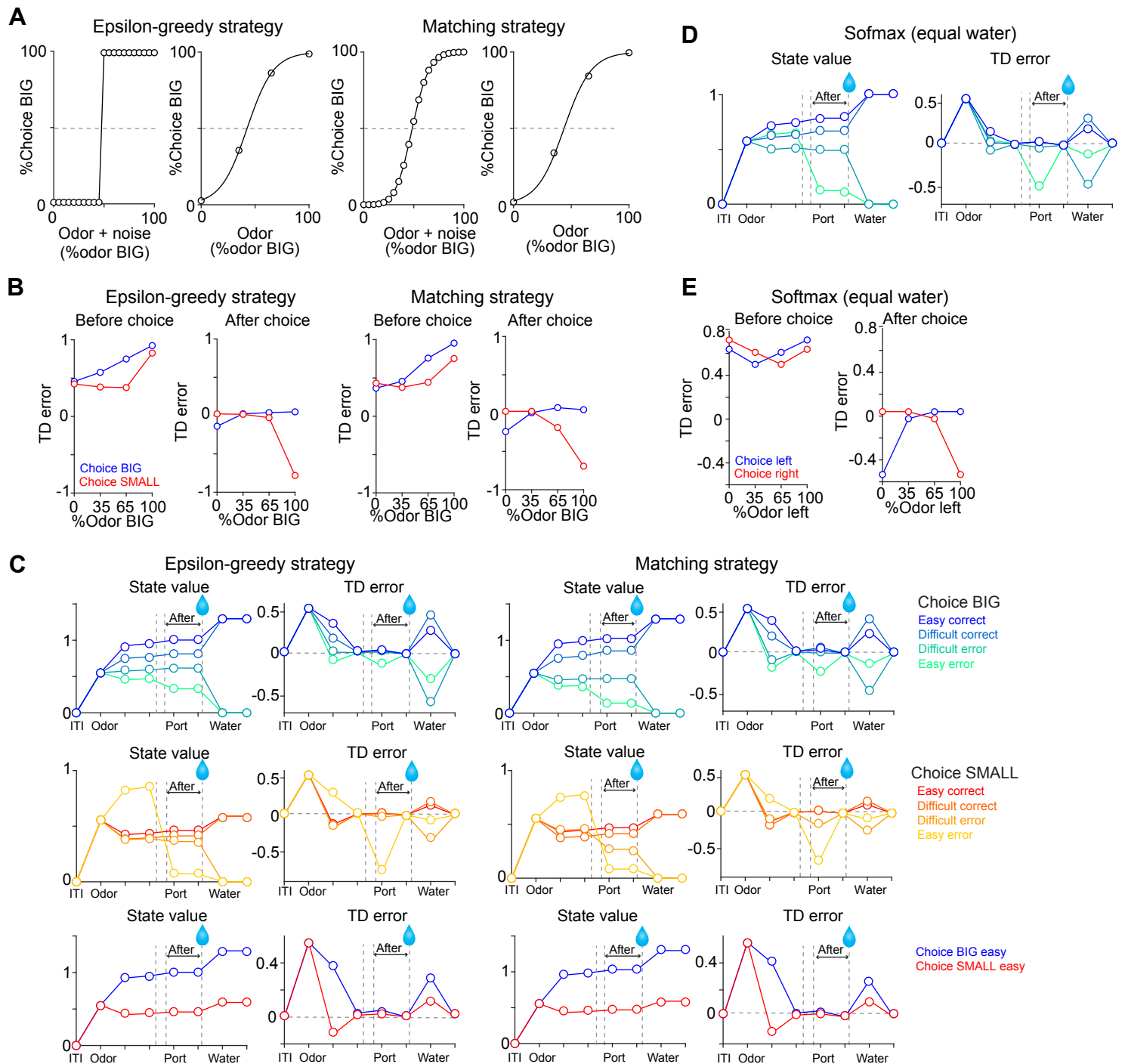
Figure S4

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.