Striatal dopamine pulses follow a temporal discounting spectrum

Wei Wei[1,†], **Ali Mohebi**[1,†], **and Joshua D. Berke**[1,2,3,4,5,†,✉]

[1]Department of Neurology,
[2]Department of Psychiatry and Behavioral Sciences,
[3]Neuroscience Graduate Program,
[4]Kavli Institute for Fundamental Neuroscience,
[5]Weill Institute for Neurosciences, University of California, San Francisco, United States
[†]Authors contributed equally

The striatum is critical for making decisions based on predictions of future reward. These predictions can be updated by brief pulses of dopamine, encoding reward prediction errors. However, it is unclear how this mechanism handles the need to generate predictions over multiple time horizons: from seconds or less (if singing a song) to potentially hours or more (if hunting for food). Here we monitor and model dopamine pulses across distinct striatal subregions, and find that these reflect predictions over distinct temporal scales. Dopamine dynamics systematically accelerated from ventral to dorsal-medial to dorsal-lateral striatum, in the tempo of their spontaneous fluctuations, their integration of prior rewards, and their discounting of future rewards. In this way parallel striatal circuits can achieve a more comprehensive set of value computations, to guide a broad range of reward-related behaviors.

Dopamine | Reward prediction error | Value |Reinforcement Learning Temporal discounting

Correspondence: *joshua.berke@ucsf.edu*

Introduction

How much should we care about the future? It makes sense to discount rewards that are far away in time - among other reasons, they are less certain to occur at all (1). Yet some worthwhile rewards take time and work to acquire. We must not discount such delayed rewards too quickly, to maintain motivation and avoid choosing less favorable, but faster, gratification. Excessive discounting - i.e., failure to maintain a sufficiently long time horizon - has been reported in a range of human psychiatric disorders (2), notably drug addiction (3).

The striatum is a key brain node for the maintenance and use of reward predictions ("values"; (4, 5). When these values are found to be inaccurate, a reward prediction error (RPE) is generated. RPEs can be encoded by brief ("phasic") increases in the firing of midbrain dopamine (DA) cells (6–10). The corresponding pulse of striatal DA release (11, 12) may engage striatal synaptic plasticity (13, 14) to update stored values.

DA RPEs have been classically considered a unitary signal that is broadcast globally across striatum and frontal cortex (6). However, a single RPE signal implies a single underlying value, with a single time horizon. A single time horizon might struggle to accommodate the range of decisions that we and other animals need to make (15). For example, during rapid production of motor sequences (e.g. birdsong) the relevant time horizon may be a fraction of a second (16) but while foraging for food an appropriate time horizon may be orders of magnitude longer (1). Evaluation using multiple discount factors in parallel can better account for behavior (17, 18) and also improve performance of artificial learning systems (19, 20).

Furthermore, there is now substantial evidence for heterogeneity of DA cell firing (9, 21), and DA release across distinct striatal subregions (12, 22–28) that are components of distinct large-scale loop circuits (29). These loops are proposed to serve as distinct levels of a hierarchical reinforcement learning architecture (30) with more dorsal/lateral striatal subregions concerned with specific motoric details while more ventral/medial areas help to organize behavior over longer time scales (31). Theoretical studies have proposed a corresponding gradient of temporal discount factors across striatum (17) and there is evidence for graded discounting from human fMRI (32). Yet how DA signals in distinct striatal subregions reflect the use of different time horizons has not been examined, to our knowledge.

We compared DA dynamics across multiple striatal subregions in both instrumental and Pavlovian tasks, with a focus on phasic RPE coding. We found systematic variation in both the tempo of spontaneous DA fluctuations and in the patterns of cue-evoked responses. We show that these patterns can be largely explained by a change in the discount rate for the underlying reward predictions, consistent with a portfolio of time horizons for decision-making.

Results

We used fiber photometry of the fluorescent DA sensor dLight1.3b (12, 34) to observe DA release fluctuations in the striatum of awake, unrestrained rats. We focused on three standard subregions (Fig. 1A): dorsolateral (DLS), dorsal-medial (DMS) and ventral (VS). These receive distinct patterns of cortical input (35) and are often considered to have distinct "motor", "cognitive" and "limbic" functions respectively (36, 37).

We first examined spontaneous DA fluctuations, unconstrained by task performance. DA signals showed clearly distinct dynamics in each subregion (Fig. 1B), changing most rapidly in DLS and most slowly in VS (Fig. 1C). When pre-

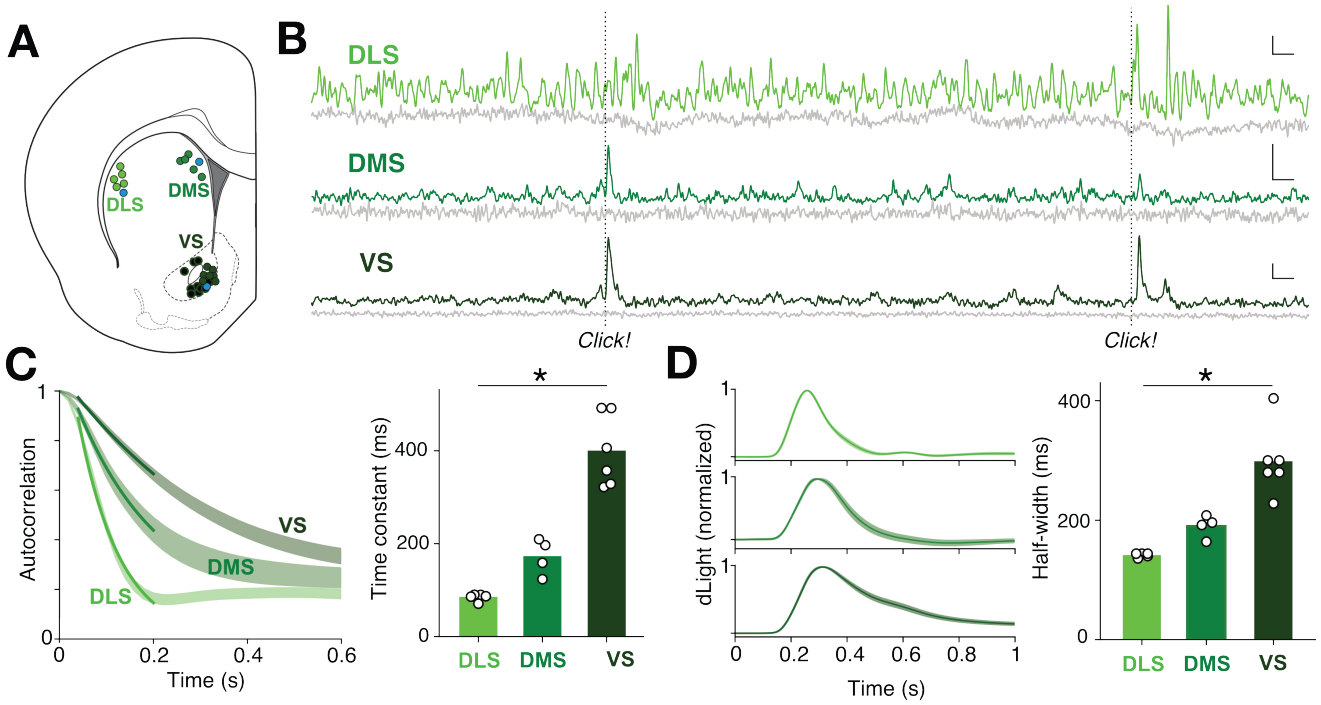


Fig. 1. Dopamine tempo depends on striatal subregion. A, A rat brain atlas section (33), showing approximate locations of fiber optic tips (circles) within striatal subregions. Blue circles indicate the locations for the recordings in B, and black-filled circles indicate locations for VS bandit task recordings (12). For further details, see Supplementary Figure 1. **B**, Example showing simultaneous, raw dLight photometry (470nm) from each subregion in an awake unrestrained rat, showing activity outside of specific task performance. Green traces indicate dopamine signals, grey traces indicate corresponding control signals (interleaved 415nm measurements). Occasional random deliveries of sugar pellet rewards are marked as "Click!" (familiar food hopper activation sound). Scale bars: 1s, 1% dF/F. Recording locations are marked by filled-in circles in A. **C**, left, Average autocorrelogram functions for spontaneous dLight signals in each subregion. Bands show $\pm$ SEM, and darker lines indicate best-fit exponential decay for the range 40ms to 200ms. Data are from n=10 rats over 15 recording sessions each; fiber placements n=5 DLS, n=4 DMS, n=6 VS). Right, decay time constant depends on subregion (ANOVA: $F(2,12) = 50.3, p = 1.5 \times 10^{-6}$). **D**, Left, average dLight signal change after an unexpected reward click; right, duration (at half maximum) of signal increase depends on subregion (ANOVA: $F(2,12) = 24.0, p = 6.5 \times 10^{-5}$).

sented with a familiar, but unexpected, reward cue - the click of a hopper dispensing a sugar pellet - all three subregions showed a transient DA response (Fig. 1D). This pulse lasted longest in VS, and was briefest in DLS (Fig. 1D). Prior studies using voltammetry and this same reward cue found DA to be evoked selectively in VS (22), but our use of dLight may have revealed DLS/DMS responses that are too brief to readily detect with voltammetry. Briefer DA signals in more dorsal regions are consistent with studies showing faster rates of DA uptake, across species (38–40), although this alone appears insufficient to explain the highly distinct spontaneous DA patterns in simultaneous recordings (Fig. 1B).

We then considered how this reward cue response is affected by recent reward history, using an instrumental "bandit" task (Fig. 2A; (12, 27). Well-trained rats made leftward or rightward nose-pokes, and as they entered the chosen side port ("Side-In") they sometimes heard the food hopper click (reward probabilities varied from 10-90%, in blocks). We previously reported that at reward delivery, both VTA DA cell firing and VS DA release scale with RPE - i.e. they are greater when reward expectation is lower, due to fewer recent trials being rewarded. We now also observed positive DA RPE coding in DLS and DMS (Fig. 2B), although the DA pulse was briefer in DMS compared to VS, and again remarkably brief in DLS (Fig. 2C; half width $135 \pm 29ms$ S.E.M.). On omission trials, DA dipped in all three subregions, and the duration of this dip was also subregion-dependent (Fig. 2D).

Despite being present in each subregion, the DA pulse was not a "global" RPE signal, since it did not reflect the same underlying value in each subregion. Expectation of future reward can reflect past reward history over a range of possible (retrospective) time horizons (41, 42). We estimated the time horizon used for the DA RPE signal, using first a time-based leaky integrator of rewards (43). This model has a single parameter τ: larger τ corresponds to a longer time horizon, allowing rewards to better summate over multiple trials (Fig. 2E). For each fiber location, we determined the τ that produced the strongest correlation between DA pulses and RPE. We observed a systematic relationship to location: best-fit τ was shortest in DLS, intermediate in DMS, and longest in VS (Fig. 2F), consistent with a time horizon spectrum. As an alternative estimate of value time horizon (8), we considered how quickly or slowly values are updated from trial-to-trial: smaller updates result in values being dependent on the outcome of a longer history of previous trials. Using a simple, trial-based delta-rule model (44) we determined the learning rate α that maximized DA: RPE correlations. Best-fit α was highest in DLS and lowest in VS (Fig. 2G), again indicating that VS is concerned with rewards integrated over more prolonged time frames.

The RPE theory of phasic DA is based in large measure on DA cell responses to Pavlovian conditioned cues that pre-

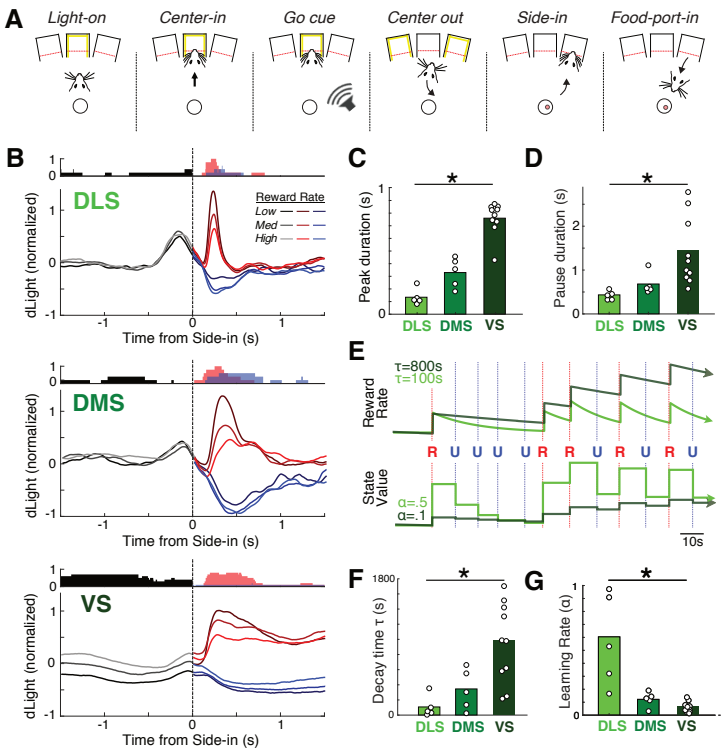


Fig. 2. Dopaminergic prediction errors depend upon subregion-specific reward history timescales. A, Event sequence in the bandit (trial-and-error) task. Rats discover if the current trial is rewarded at Side-In, when the food hopper may (or may not) click. **B**, Mean dLight DA signals aligned on the Side-In event. Data for DLS and DMS were recorded simultaneously from n=5 rats, each with one fiber in each target. VS data is from n=10 fibers in 7 distinct rats. Signals are broken down by recent reward rate (in terciles), with higher reward rate in brighter colors. After Side-In, signals are further broken down by rewarded (red) and unrewarded (blue) trials. Histogram above each plot shows the fraction of signals that significantly depended on reward rate (linear regression, p<0.01), consistent (after Side-In) with RPE coding. Reward rates were calculated using a leaky integrator of reward receipts (see Methods and E below), choosing the tau parameter for each subregion separately to maximize RPE coding (fits to behavior or alternative models of reward prediction gave similar results, Supplementary Fig. 2). The bump before Side-In (most prominent for DLS) is the response to the Go cue, smeared by variability in reaction and movement times. **C**, The duration of the DA peak significantly varies by subregion (one-way ANOVA, $F(2,18) = 25.2, p = 8.23 \times 10^{-6}$; measured at half-maximum on rewarded trials in the 1s period after Side-In). **D**, as C but for DA pause duration (one-way ANOVA, $F(2,18) = 5.64, p = 0.014$; unrewarded trials, half-minimum, 4s after Side-In). **E**, Top, illustration of leaky integrator estimation of reward rate, for an example sequence of trials (R = rewarded, U = unrewarded) and the tau decay parameter set to either 100 or 800s. Bottom, estimating reward expectation for the same example sequence using a simple delta-rule model, with one update per trial and learning rate parameter set to either 0.1 or 0.5. **F**, The leaky-integrator τ that maximizes correlation between RPE and DA after Side-in significantly varies by subregion (one-way ANOVA, F(2,18) = 7.99, p = 0.0039). **G**, The delta-rule learning rate α that maximizes correlation between RPE and DA after Side-in significantly varies by subregion (one-way ANOVA, $F(2,18) = 11.62, p = 0.0007$). The strongest correlations are seen in DLS with a shorter time horizon (small τ, or large α) and in VS with a longer time horizon (large τ, or small α).

dict future rewards (6, 10). Such responses are diminished when the rewards are more distant, in a manner consistent with temporal discounting (45, 46). We therefore examined DA cue responses in a Pavlovian approach task (Fig. 3A). Auditory cues (trains of 2, 5, 9 kHz tone pips) predicted sugar pellet delivery a few seconds later with distinct probabilities (75, 25, 0%; see Methods). Each trial presented one of the cues, or an uncued reward delivery, in random order, with a 15-30s delay between trials. Rats were trained for 15 days (each day had 60 trials of each type). Early on, all cues increased the likelihood of food hopper entry (Fig. 3B), consistent with generalization between cues (47). However, over the course of training (3600 trials total) rats showed increasing discrimination, entering the food hopper in proportion to cued reward probability (Fig. 3B).

Strikingly-different phasic DA patterns were seen for each subregion (Fig. 3C,D; Supplementary. Fig. 3). In well-trained rats, DMS DA showed strong RPE coding at cue onset (Fig. 3C). Specifically, the 75% cue produced a strong DA pulse, the 25% cue a much smaller pulse, and the 0% cue a transient dip in DA. VS cue responses also scaled with RPE but showed worse discrimination between cues, particularly on early training days, and remained positive for all cues throughout training. Concordant results of VS DA increases to a learned 0% cue (CS-) have been previously reported, and attributed to generalization between cues (6, 48). Finally, in DLS the predictive cues evoked much smaller DA responses (relative to unpredicted reward delivery). This did not simply reflect a failure of DLS-related circuits to learn: the DLS DA pulse at reward delivery was substantially diminished if preceded by the 75% cue (Fig. 3C), consistent with an acquired reward prediction.

We reasoned that these subregional differences could reflect distinct time horizons for value computations. If future rewards are discounted especially fast in DLS-related circuits, even a brief delay would substantially diminish the value indicated by cues (Fig. 4A). To assess this idea we turned to computational models that address the evolution of value within trials. We first applied a standard, simple model in which the cue-reward interval is divided into a regular sequence of sub-states (the "complete serial compound", CSC; (49). Over the course of learning, value propagates backwards along the sub-state chain (50). As expected, when we compared model versions with distinct discount rates (γ), rapid discounting reproduced the DLS pattern of smaller cue responses (Fig. 4B-D) despite a cue-dependent response to reward delivery (Fig. 4B). Including overlap between cue representations allowed the CSC to also reproduce the generalization between cues early in training (Fig. 4D).

However, our CSC model of the cue-reward interval could not readily account for the slower, poorer cue discrimination in VS (Fig. 4C), and is incapable of reproducing the negative response to the 0% cue we saw in DMS. This model is not designed to handle prolonged time horizons that might span multiple trials (15). Furthermore, the splitting of experience into discrete, equally-fine sub-states becomes ever more artificial as inter-trial intervals get larger and more variable (51, 52).

We therefore turned to an alternative approach for estimating the evolution of values, using recurrent neural net-

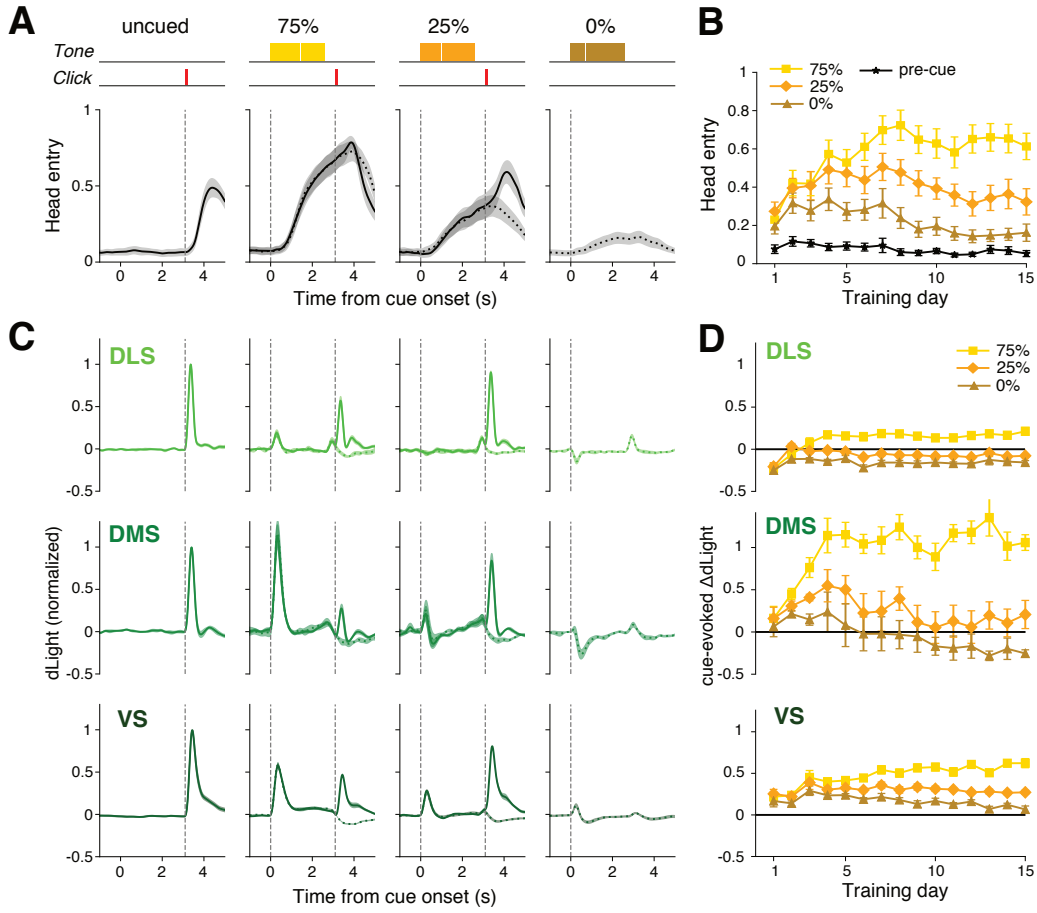


Fig. 3. Subregion-specific dopamine responses to reward-predictive cues. A, Top, the Pavlovian task consists of four trial types, selected at random, with differing reward probabilities. Bottom, after training cues increase anticipatory head entries into the reward port (fraction of trials, mean ± SEM), and this scales with reward probability. Data shown are averages from training days 13-15, for n = 10 rats. **B**, During early training days rats increase their behavioral responses to all cues, before progressively learning to discriminate between cues (error bars: SEM). Points show average head entry over a 0.5 s epoch just before cue onset (black) or just after cue offset (colors; i.e. immediately before the time that reward could be delivered). **C**, Average dLight signal change for each trial type after training (days 13-15, n = 10 rats with fibers in DLS (n = 5), DMS (n = 4) and VS (n = 7). Solid lines = rewarded trials, dotted lines = unrewarded. **D**, Time course of dopamine increases to each cue in each subregion over training (mean ± SEM). For responses on individual trials on the first day of training, see Supplementary Figure 3.

works (RNNs; (53, 54). In our RNN model (Fig. 5B; see Methods), multiple sub-networks each use reinforcement learning to generate distinct values in tandem (55), but with distinct discount factors (56). The model has no discrete states and time is not explicitly represented, but rather is implicit within network population dynamics (57). With the simple assumption that time horizon increases from DLS to DMS to VS, the RPEs generated by the model recapitulated key features of the rat striatal DA pulses (Fig. 5C,D). These include the diminutive DLS responses as before, but also the negative DMS response to the 0% cue, and poor VS cue discrimination compared to DMS (especially earlier in training).

With extended RNN training the "DLS" and "DMS" responses to cues remained relatively stable, but "VS" cue discrimination continued to improve, eventually also acquiring negative RPE responses to the 0% cue (Fig. 5; Supplementary. Fig. 4). In other words, a discount factor very close to 1 made learning slow, consistent with prior observations in reinforcement learning (58). With hindsight, this made intuitive sense. If a time horizon encompasses many trials, it will include multiple rewards regardless of which cue is presented (Fig. 5A). Discriminating between cues is therefore harder, and slower to learn. By contrast, if the time horizon for DMS is on the order of one trial, the average outcomes following distinct cues are very different (closer to the nominal 75, 25,%) and so learning the distinct associated values is much less challenging.

The idea of distinct time horizons thus provides a concise explanation for the subregional differences in Pavlovian cue-evoked DA pulses. DLS responses are weaker because the cues indicate reward that is too far away in time, given a short time horizon. VS responses are slower to discriminate, because the rewards that follow each cue are not very different, over a long time horizon. And DMS shows stronger, well-discriminating responses because its intermediate time horizon best matches the actual time scale of predictions provided by the Pavlovian cues.

Discussion

Our overall conclusion is that, across a range of behavioral contexts, DA pulses in distinct striatal subregions reveal a spectrum of underlying time scales for evaluation. This supports the proposal that parallel cortical-basal ganglia circuits organize behavior over distinct time scales (30, 31, 59, 60) that are also apparent in the dynamics of single neuron firing (61, 62). DLS microcircuits appear to be specialized for faster information processing, with features including faster DA reuptake and a higher proportion of fast-spiking interneurons to dictate fine timing (63). Conversely, more prolonged representations in VS circuits (64) may be important to achieve learning over longer time horizons (59). Some

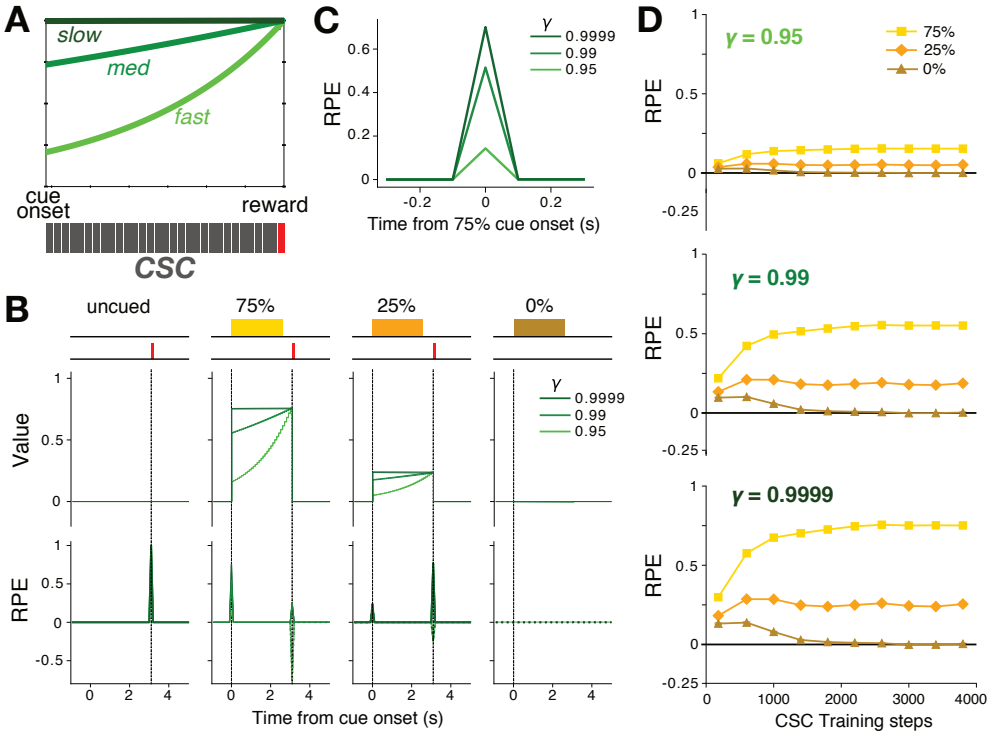


Fig. 4. Fast temporal discounting can explain weaker DLS cue responses. A, Top, faster temporal discounting erodes the value indicated by the onset of a reward-predictive cue, even if the reward is certain to appear. Bottom, in the CSC model the cue-reward interval is divided into a fixed set of brief sub-states (we used 100ms duration). **B**, Values, and corresponding temporal-difference RPEs, for the CSC model after training in the Pavlovian task (step 3800). Discount factor γ was set to 0.95 (light green, "fast"), 0.99 (mid-green), or 0.9999 (dark green, "slow"). **C**,Close-up of the CSC RPE response to the 75% cue. Even if the cued reward probability is high (75%), RPEs at cue onset are weaker when the discount factor is lower (RPEs at reward delivery are unchanged). **D**, Development of RPEs at cue onsets with training. Note that cue discrimination after training is larger if γ is closer to 1 (plotted in more detail in Supplementary Fig. 4). Overlapping cue representations cause this CSC model to produce a positive RPE to the 0% cue early in training, but this eventually fades to zero.

studies using fMRI have suggested that VS circuits discount especially rapidly (32, 65) and may therefore promote impulsive behavior. By contrast, our results are consistent with a large body of literature demonstrating a critical role for VS in avoiding maladaptive impulsive behavior (66, 67), by instead promoting work to obtain delayed rewards (68, 69).

We used a standard systems neuroscience approach: a behavioral session with many individual trials, and cues that are meaningful for that specific trial. But our results emphasize that animals, and their neural sub-circuits, do not necessarily process information in a corresponding trial-based manner (70). The notion that VS-related circuits use a time horizon for reward that can span many trials may explain other, previously-puzzling observations. In particular, voltammetry studies have observed especially large VS DA transients as each session begins (e.g. (71). This makes sense if - from the VS perspective - the onset of the first trial indicates that the animal is likely to receive multiple rewards "soon", across multiple trials.

However, some features of our signals remain to be accounted for. First, our novel auditory cues simultaneously evoked a positive DA response in VS, and a negative response in DLS (Fig. 3D), even on the very first presentation (Supp. Fig. 3; for distinct results, see (26)). DA increases to novel cues have been previously ascribed to a "novelty bonus" (72), which can accelerate learning(73). A novelty bonus may be especially helpful for VS when facing the challenging problem of learning values over longer time horizons. Why DLS should instead show a "novelty penalty" is less clear, but may reflect DLS involvement in motor learning with forward models (74). If DLS-related circuits are comparing the sensory feedback from actions to an intended template (such as a tutor bird's song), then a novel cue would likely be classified as worse-than-expected (i.e. a negative RPE; (16).

Second, striatal subregions can differently care about other aspects of behavior. As some examples: DA in DLS responds relatively more strongly to Go cues (Fig. 2B); DA in DMS seems to be preferentially engaged with contralateral movements (24, 75); and DA in VS preferentially scales with value during approach behaviors (12, 27, 76). Third, the magnitude of learned DMS response to cues was larger (and more variable) than we would expect for an RPE with intermediate time horizon (Fig. 3C) - indeed sometimes exceeding the response to unpredicted reward. Each of these findings merits further investigation.

Our RNN model describes how temporal discount factors sculpt values (and hence RPEs), but does not provide a detailed mechanistic account of how cortical-basal ganglia loop circuits manage the passage of time and learning over delays. Learning with short time intervals (subsecond-seconds) can be supported by eligibility traces (14), sequential activity patterns (77) and potentially interactions with cerebellum (78). Longer delays to reward pose a more challenging problem of credit assignment, that may be mitigated by interactions with the hippocampus and off-line replay of experiences (79).

Using multiple sub-agents with distinct discount factors may be an optimal strategy in a complex and changing environment (42, 80). However, this creates the challenge of how to appropriately integrate multiple, conflicting predictions (81), and can drive apparently irrational behavior. Even if individual sub-agents simply discount future rewards at a

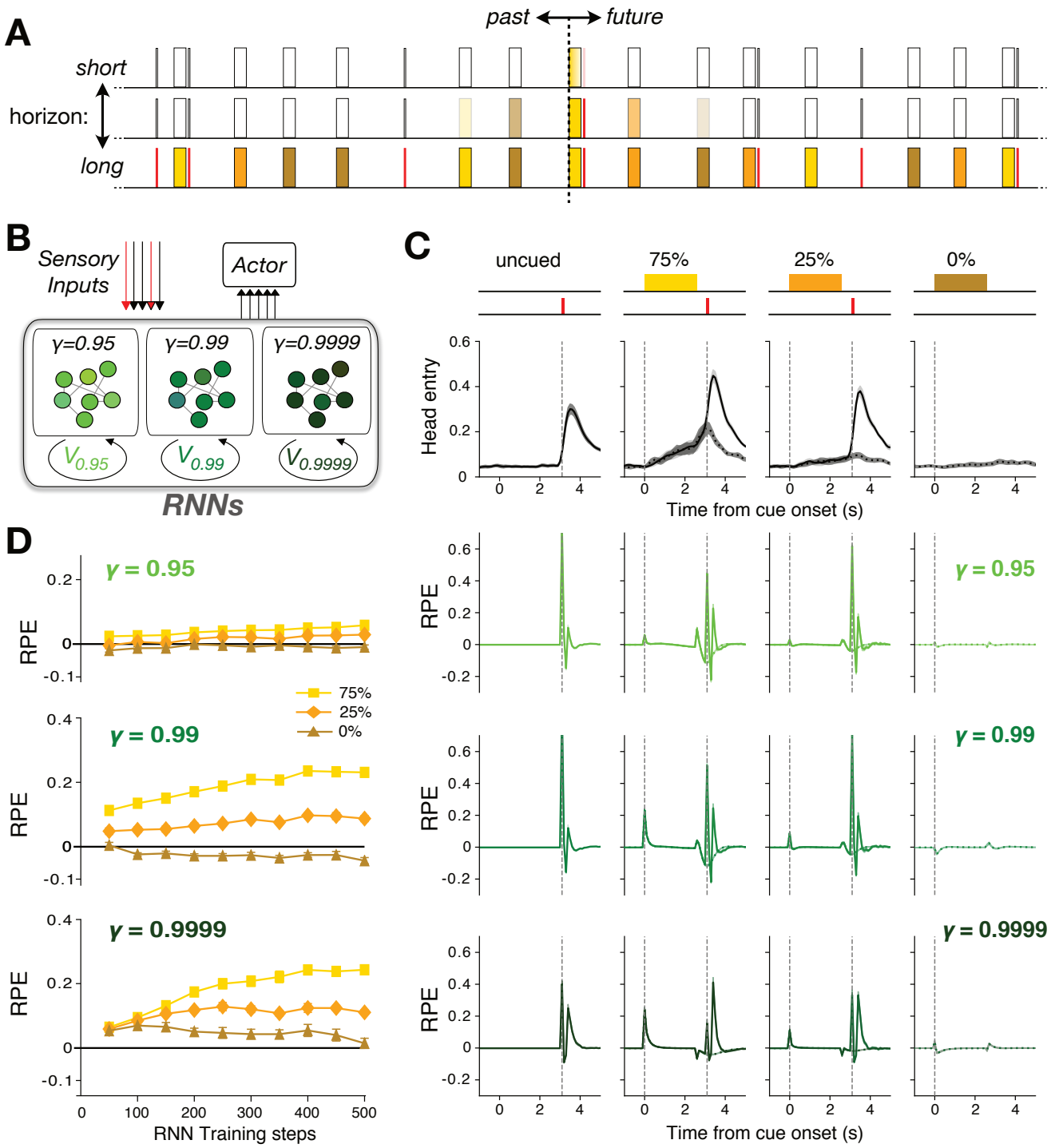


Fig. 5. Long time horizons can explain slow VS cue discrimination. A, Schematic of part of a long random sequence of trials within a single training session, with colors indicating the cue in each trial. At any given moment, a reinforcement learning agent may be estimating the amount of reward that is coming "soon", and updating such estimates based on what happened "recently". If the time horizon is long, "soon" can encompass rewards likely to occur across multiple trials, even if the current trial has 0% chance of reward. **B**, Schematic of recurrent neural network model, with three distinct pools of LSTM units. Each pool receives the same sensory inputs, but maintains its own value output based on a distinct discount factor γ (0.95, 0.99, or 0.9999, all with a 100ms timestep). All three pools project to the Actor, which generates the probability of nose-poking. **C**, Model poke probability (top) and temporal-difference RPEs for each LSTM pool, after 500 training steps. **D**, Development of RPEs at cue onsets across training (see Supplementary. Fig. 4 for extended training).

constant rate (exponential discounting), collectively they can display a declining discount rate (approximately hyperbolic; (17). This inconsistency is a feature of animal and human economic behavior: choices show increasing impatience as rewards become more imminent (82, 83). Such inconsistencies may be an unavoidable price we have to pay, in exchange for a discounting spectrum that enables us to efficiently learn adaptive interactions with a complex world.

ACKNOWLEDGEMENTS

We thank Vijay Namboodiri, Robert Schmidt, and members of the Berke Lab for their comments on a prior version of the manuscript. Support for the work presented here was provided by NIDA, NINDS, NIMH, and UCSF.

Bibliography

1. David W Stephens and Dack Anderson. The adaptive value of preference for immediacy: when shortsighted rules have farsighted consequences. *Behavioral Ecology*, 12(3):330–339, 2001.
2. Michael Amlung, Emma Marsden, Katherine Holshausen, Vanessa Morris, Herry Patel, Lana Vedelago, Katherine R Naish, Derek D Reed, and Randi E McCabe. Delay discounting as a transdiagnostic process in psychiatric disorders: a meta-analysis. *JAMA Psychiatry*, 76(11):1176–1186, 2019.
3. Warren K Bickel and Lisa A Marsch. Toward a behavioral economic understanding of drug dependence: delay discounting processes. *Addiction*, 96(1):73–86, 2001.
4. Kazuyuki Samejima, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. Representation of action-specific reward values in the striatum. *Science*, 310(5752):1337–1340, 2005.
5. Joseph W Kable and Paul W Glimcher. The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633, 2007.
6. Wolfram Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1–27, 1998.
7. Genela Morris, David Arkadir, Alon Nevet, Eilon Vaadia, and Hagai Bergman. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron*, 43(1):133–143, 2004.
8. Hannah M Bayer and Paul W Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141, 2005.
9. Ethan S Bromberg-Martin, Masayuki Matsumoto, Hiroyuki Nakahara, and Okihide Hikosaka. Multiple timescales of memory in lateral habenula and dopamine neurons. *Neuron*, 67(3): 499–510, 2010.
10. Jeremiah Y Cohen, Sebastian Haesler, Linh Vong, Bradford B Lowell, and Naoshige Uchida. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, 482(7383):85–88, 2012.
11. Andrew S Hart, Robb B Rutledge, Paul W Glimcher, and Paul EM Phillips. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience*, 34(3):698–704, 2014.
12. Ali Mohebi, Jeffrey R Pettibone, Arif A Hamid, Jenny-Marie T Wong, Leah T Vinson, Tommaso Patriarchi, Lin Tian, Robert T Kennedy, and Joshua D Berke. Dissociable dopamine dynamics for learning and motivation. *Nature*, 570(7759):65–70, 2019.
13. John NJ Reynolds, Brian I Hyland, and Jeffery R Wickens. A cellular mechanism of reward-related learning. *Nature*, 413(6851):67–70, 2001.
14. Sho Yagishita, Akiko Hayashi-Takagi, Graham CR Ellis-Davies, Hidetoshi Urakubo, Shin Ishii, and Haruo Kasai. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620, 2014.
15. Nathaniel D Daw and David S Touretzky. Long-term reward prediction in td models of the dopamine system. *Neural Computation*, 14(11):2567–2583, 2002.
16. Vikram Gadagkar, Pavel A Puzerey, Ruidong Chen, Eliza Baird-Daniel, Alexander R Farhang, and Jesse H Goldberg. Dopamine neurons encode performance error in singing birds. *Science*, 354(6317):1278–1282, 2016.
17. Zeb Kurth-Nelson and A David Redish. Temporal-difference reinforcement learning with distributed representations. *PLoS One*, 4(10):e7362, 2009.
18. Gary A Kane, Aaron M Bornstein, Amitai Shenhav, Robert C Wilson, Nathaniel D Daw, and Jonathan D Cohen. Rats exhibit similar biases in foraging and intertemporal choice tasks. *eLife*, 8:e48429, 2019.
19. Chris Reinke, Eiji Uchibe, and Kenji Doya. Average reward optimization with multiple discounting reinforcement learners. In *International Conference on Neural Information Processing*, pages 789–800. Springer, 2017.
20. William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.
21. Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.
22. Holden D Brown, James E McCutcheon, Jackson J Cone, Michael E Ragozzino, and Mitchell F Roitman. Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. *European Journal of Neuroscience*, 34(12):1997–2006, 2011.
23. Ingo Willuhn, Lauren M Burgeno, Barry J Everitt, and Paul EM Phillips. Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. *Proceedings of the National Academy of Sciences*, 109(50):20703–20708, 2012.
24. Nathan F Parker, Courtney M Cameron, Joshua P Taliaferro, Junuk Lee, Jung Yoon Choi, Thomas J Davidson, Nathaniel D Daw, and Ilana B Witten. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience*, 19(6):845–854, 2016.
25. Mark W Howe and Daniel A Dombeck. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature*, 535(7613):505–510, 2016.
26. William Menegas, Benedicte M Babayan, Naoshige Uchida, and Mitsuko Watabe-Uchida. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife*, 6:e21886, 2017.
27. Arif A Hamid, Jeffrey R Pettibone, Omar S Mabrouk, Vaughn L Hetrick, Robert Schmidt, Caitlin M Vander Weele, Robert T Kennedy, Brandon J Aragona, and Joshua D Berke. Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, 19(1):117–126, 2016.
28. Iku Tsutsui-Kimura, Hideyuki Matsumoto, Korleki Akiti, Melissa M Yamada, Naoshige Uchida, and Mitsuko Watabe-Uchida. Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. *eLife*, 9:e62390, 2020.
29. Garrett E Alexander and Michael D Crutcher. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, 13(7):266–271, 1990.
30. Michael J Frank and David Badre. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral Cortex*, 22(3):509–526, 2012.
31. Makoto Ito and Kenji Doya. Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, 21(3):368–373, 2011.
32. Saori C Tanaka, Kenji Doya, Go Okada, Kazutaka Ueda, Yasumasa Okamoto, and Shigeto Yamawaki. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7(8):887–893, 2004.
33. George Paxinos and Charles Watson. *The rat brain in stereotaxic coordinates: hard cover edition*. Elsevier, 2007.
34. Tommaso Patriarchi, Jounhong Ryan Cho, Katharina Merten, Mark W Howe, Aaron Marley, Wei-Hong Xiong, Robert W Folk, Gerard Joey Broussard, Ruqiang Liang, Min Jee Jang, et al. Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science*, 360(6396), 2018.
35. Barbara J Hunnicutt, Bart C Jongbloets, William T Birdsong, Katrina J Gertz, Haining Zhong, and Tianyi Mao. A comprehensive excitatory input map of the striatum reveals novel functional organization. *eLife*, 5:e19103, 2016.
36. Pieter Voorn, Louk JMJ Vanderschuren, Henk J Groenewegen, Trevor W Robbins, and Cyriel MA Pennartz. Putting a spin on the dorsal–ventral divide of the striatum. *Trends in Neurosciences*, 27(8):468–474, 2004.
37. Bryan D Devan, Nancy S Hong, and Robert J McDonald. Parallel associative processing in the dorsal striatum: segregation of stimulus–response and cognitive control subregions. *Neurobiology of Learning and Memory*, 96(2):95–120, 2011.
38. Sara R Jones, Paul A Garris, Clinton D Kilts, and R Mark Wightman. Comparison of dopamine uptake in the basolateral amygdaloid nucleus, caudate-putamen, and nucleus accumbens of the rat. *Journal of Neurochemistry*, 64(6):2581–2589, 1995.
39. Stephanie J Cragg, Christopher J Hille, and Susan A Greenfield. Functional domains in dorsal striatum of the nonhuman primate are defined by the dynamic behavior of dopamine. *Journal of Neuroscience*, 22(13):5705–5712, 2002.
40. Erin S Calipari, Kimberly N Huggins, Tiffany A Mathews, and Sara R Jones. Conserved dorsal–ventral gradient of dopamine release and uptake rate in mice, rats and rhesus macaques. *Neurochemistry International*, 61(7):986–991, 2012.
41. Alberto Bernacchia, Hyojung Seo, Daeyeol Lee, and Xiao-Jing Wang. A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience*, 14(3):366–372, 2011.
42. Kiyohito Iigaya, Yashar Ahmadian, Leo P Sugrue, Greg S Corrado, Yonatan Loewenstein, William T Newsome, and Stefano Fusi. Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. *Nature Communications*, 10(1):1–14, 2019.
43. Leo P Sugrue, Greg S Corrado, and William T Newsome. Matching behavior and the representation of value in the parietal cortex. *Science*, 304(5678):1782–1787, 2004.
44. Sangil Lee, Joshua I Gold, and Joseph W Kable. The human as delta-rule learner. *Decision*, 7(1):55, 2020.
45. Shunsuke Kobayashi and Wolfram Schultz. Influence of reward delays on responses of dopamine neurons. *Journal of Neuroscience*, 28(31):7837–7846, 2008.
46. Kazuki Enomoto, Naoyuki Matsumoto, Hitoshi Inokawa, Minoru Kimura, and Hiroshi Yamada. Topographic distinction in long-term value signals between presumed dopamine neurons and presumed striatal projection neurons in behaving monkeys. *Scientific Reports*, 10(1):1–14, 2020.
47. Robert Colin Honey. Stimulus generalization as a function of stimulus novelty and familiarity in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, 16(2):178, 1990.
48. Jeremy J Day, Mitchell F Roitman, R Mark Wightman, and Regina M Carelli. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature Neuroscience*, 10(8):1020–1028, 2007.
49. Richard S Sutton and Andrew G Barto. Time-derivative models of pavlovian reinforcement. In M. Gabriel and J. Moore, editors, *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, chapter 12, pages 497–537. MIT Press, 1990.
50. P Read Montague, Peter Dayan, and Terrence J Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16(5): 1936–1947, 1996.
51. Elliot A Ludvig, Richard S Sutton, and E James Kehoe. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, 20(12):3034–3054, 2008.
52. Vijay Mohan K Namboodiri. What is the state space of the world for real animals? *bioRxiv*, 2021.
53. H Francis Song, Guangyu R Yang, and Xiao-Jing Wang. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 6:e21492, 2017.
54. Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-

reinforcement learning system. *Nature Neuroscience*, 21(6):860–868, 2018.
55. Kenji Doya, Kazuyuki Samejima, Kenichi Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural Computation*, 14(6):1347–1369, 2002.
56. A David Redish. Addiction as a computational process gone awry. *Science*, 306(5703): 1944–1947, 2004.
57. Uma R Karmarkar and Dean V Buonomano. Timing in the absence of clocks: encoding time in neural network states. *Neuron*, 53(3):427–438, 2007.
58. Vektor Dewanto and Marcus Gallagher. Examining average and discounted reward optimality criteria in reinforcement learning. *arXiv preprint arXiv:2107.01348*, 2021.
59. Masahiko Haruno and Mitsuo Kawato. Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fmri examination in stimulus-action-reward association learning. *Neural Networks*, 19(8):1242–1254, 2006.
60. Matthew M Botvinick, Yael Niv, and Andew G Barto. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.
61. John D Murray, Alberto Bernacchia, David J Freedman, Ranulfo Romo, Jonathan D Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, et al. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17(12): 1661–1663, 2014.
62. Makoto Ito and Kenji Doya. Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed-and free-choice tasks. *Journal of Neuroscience*, 35(8):3499–3514, 2015.
63. Joshua D Berke. Functional properties of striatal fast-spiking interneurons. *Frontiers in Systems Neuroscience*, 5:45, 2011.
64. Alexxai V Kravitz, David E Moorman, Alexis Simpson, and Laura L Peoples. Session-long modulations of accumbal firing during sucrose-reinforced operant behavior. *Synapse*, 60 (6):420–428, 2006.
65. Samuel M McClure, David I Laibson, George Loewenstein, and Jonathan D Cohen. Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695): 503–507, 2004.
66. Rudolf N Cardinal, David R Pennicott, C Lakmali, Sugathapala, Trevor W Robbins, and Barry J Everitt. Impulsive choice induced in rats by lesions of the nucleus accumbens core. *Science*, 292(5526):2499–2501, 2001.
67. Jeffrey W Dalley, Tim D Fryer, Laurent Brichard, Emma SJ Robinson, David EH Theobald, Kristjan Lääne, Yolanda Peña, Emily R Murphy, Yasmene Shah, Katrin Probst, et al. Nucleus accumbens d2/3 receptors predict trait impulsivity and cocaine reinforcement. *Science*, 315 (5816):1267–1270, 2007.
68. John D Salamone and Mercè Correa. The mysterious motivational functions of mesolimbic dopamine. *Neuron*, 76(3):470–485, 2012.
69. Joshua D Berke. What does dopamine mean? *Nature Neuroscience*, 21(6):787–793, 2018.
70. Charles R Gallistel, Andrew R Craig, and Timothy A Shahan. Temporal contingency. *Behavioural Processes*, 101:89–96, 2014.
71. Matthew J Wanat, Camelia M Kuhnen, and Paul EM Phillips. Delays conferred by escalating costs modulate dopamine release to rewards but not their predictors. *Journal of Neuroscience*, 30(36):12020–12027, 2010.
72. Sham Kakade and Peter Dayan. Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6):549–559, 2002.
73. Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
74. Mitsuo Kawato and Kazuyuki Samejima. Efficient reinforcement learning: computational theories, neuroscience and robotics. *Current Opinion in Neurobiology*, 17(2):205–212, 2007.
75. Morgane M Moss, Peter Zatka-Haas, Kenneth D Harris, Matteo Carandini, and Armin Lak. Dopamine axons in dorsal striatum encode contralateral visual stimuli and choices. *Journal of Neuroscience*, 41(34):7197–7205, 2021.
76. Mark W Howe, Patrick L Tierney, Stefan G Sandberg, Paul EM Phillips, and Ann M Graybiel. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, 500(7464):575–579, 2013.
77. Shanglin Zhou, Sotiris C Masmanidis, and Dean V Buonomano. Neural sequences as an optimal dynamical regime for the readout of time. *Neuron*, 108(4):651–658, 2020.
78. Daniele Caligiore, Giovanni Pezzulo, Gianluca Baldassarre, Andreea C Bostan, Peter L Strick, Kenji Doya, Rick C Helmich, Michiel Dirkx, James Houk, Henrik Jörntell, et al. Consensus paper: towards a systems-level view of cerebellar function: the interplay between cerebellum, basal ganglia, and cortex. *The Cerebellum*, 16(1):203–229, 2017.
79. David J Foster and Matthew A Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, 2006.
80. David Meder, Nils Kolling, Lennart Verhagen, Marco K Wittmann, Jacqueline Scholl, Kristoffer H Madsen, Oliver J Hulme, Timothy EJ Behrens, and Matthew FS Rushworth. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nature Communications*, 8(1):1–11, 2017.
81. Christopher P Chambers and Federico Echenique. On multiple discount rates. *Econometrica*, 86(4):1325–1346, 2018.
82. David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.
83. George Ainslie. *Breakdown of will*. Cambridge University Press, 2001.
84. Thomas Akam and Mark E Walton. pyphotometry: Open source python based hardware and software for fiber photometry data acquisition. *Scientific Reports*, 9(1):1–11, 2019.
85. Talia N Lerner, Carrie Shilyansky, Thomas J Davidson, Kathryn E Evans, Kevin T Beier, Kelly A Zalocusky, Ailey K Crow, Robert C Malenka, Liqun Luo, Raju Tomer, et al. Intact-brain analyses reveal distinct information carried by snc dopamine subcircuits. *Cell*, 162(3): 635–647, 2015.
86. Ekaterina Martianova, Sage Aronson, and Christophe D Proulx. Multi-fiber photometry to record neural activity in freely-moving animals. *JoVE (Journal of Visualized Experiments)*, (152):e60278, 2019.
87. Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937. PMLR, 2016.
88. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
89. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, 2017.
90. John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2016.
91. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015.

Methods

Animals and Behavior. All animal procedures were approved by the University of California, San Francisco Animal Care Committee. Rats were adult wild-type Long-Evans males, bred in-house, maintained on a reverse 12:12 light : dark cycle and tested during the dark phase. All recordings were performed in an operant chamber (Med Associates), and details on both behavioral tasks have been published previously (12, 27). For the Pavlovian task each cue tone (2, 5 or 9kHz) was presented as a train of pips (100 ms on, 50 ms off) for a total duration of 2.6 s followed by a delay period of 500 ms. Trials with one of the three cues, or an unpredicted reward delivery, were delivered in pseudorandom order with a variable inter-trial interval (15–30 s, uniform distribution). Bandit task sessions used the following parameters: left–right reward probabilities were (independently-varying, randomly-selected) 10, 50 or 90% for blocks of 35-45 trials; hold period before the Go cue was 500–1,500 ms (uniform distribution). The mean number of trials was 315 (range: 216-407).

Virus and Photometry. We used a viral approach to express the genetically-encoded optical DA sensor dLight1.3b (34). Under isoflurane anesthesia, 1 µl of AAV9-CAG-dLight1.3b (2×10^{12} viral genomes per ml; Vigene) was slowly (100 nl/min) injected (Nanoject III, Drummond) through a glass micropipette targetting multiple striatal subregions: ventral (AP: 1.7, ML: 1.7, DV: 7.0 mm relative to bregma), dorsomedial (AP: 1.5, ML: 1.8, DV: -4.3) and dorsolateral (AP: 0.84, ML:3.8, DV:-4.0). During the same surgery optical fibers (400 µm core, 430 µm total diameter) attached to a metal ferrule (Doric) were inserted (target depth 200 µm higher than AAV) and cemented in place. Data were collected >3 weeks later, to allow for dLight expression. For dLight excitation blue (470 nm) and violet (405 nm; isosbestic control) LEDs were alternately switched on and off in 10ms frames: 4ms on and 6ms off (84). Excitation power at the fiber tip was set to 30 µW for each wavelength. Both excitation and emission signals passed through minicube filters (Doric) and bulk fluorescence was measured with a femtowatt detector (Newport, Model 2151) sampling at 10 kHz. Time-division multiplexing produced separate 470 nm (DA) and 405 nm (control) signals, which were then rescaled to each other via a least-square fit (85). For simultaneous recording of three areas we used a Neurophotometrics system (86); technical details were very similar except that the control wavelength was 415nm and detection was camera-based, sampling at 1kHz. Fractional fluorescence signal (dF/F) was then defined as (470–control$_{fit}$)/control$_{fit}$. For all analyses this signal was downsampled to 30 Hz and smoothed with a five-point median filter. For each Pavlovian recording session DA activity was normalized to the mean peak uncued click response in that session. We removed from analyses 3 fiber placements that produced consistently weak signals (2 DMS, 1 DLS), and we also excluded individual sessions for which the peak response was less than one standard deviation ($Z < 1$; 3 of 270 sessions excluded, 2 VS, 1 DMS). For autocorrelation analysis, we had an additional inclusion criterion of minimum dLight signal strength. This was defined as an average (across sessions, post isosbestic correction) standard deviation $Z > 0.5\%$ dF/F, measured in a window -10 to -1s before cue onsets, and resulted in one VS fiber placement being excluded. DA activity at cue time was estimated as the maximum or the minimum within a half second window after cue onset, whichever had the larger absolute value; results were not substantially different if we instead used average DA in this window (data not shown).

Histological confirmation. To verify probe placement post-mortem, animals were perfused transcardially with PBS and then 4% PFA. Implants were taken out and brains were extracted and postfixed in 4% PFA for 24 h, then placed in 30% sucrose in PBS for >48 h and sectioned at a 100µm thickness with a microtome. We used immunofluorescence staining to visualize dLight expression. Brain sections with probe placement were identified, blocked in a 0.4% Triton X-100 solution with 5% normal goat serum (NGS) for 1 h at room temperature, followed by an overnight incubation in a rabbit anti-GFP primary antibody solution (1:1000; abcam, ab290) in PBS in a cold room. Sections were washed three times in PBS for 10 min at room temperature and incubated in an Alexa 488-conjugated goat anti-rabbit secondary antibody solution (1:250) in PBS for 1 h at room temperature. Finally, sections were washed six times in PBS for 5 minutes at room temperature and then mounted onto glass slides and coverslipped using Fluoromount-G™ Mounting Medium, with DAPI. Fluorescent images were taken using a fluorescence microscope (Keyence BZ-X710) with a 2x objective lens. Fiber tip locations from both hemispheres were projected onto the same side in atlas space.

Computational Models.

Trial-level models. For the bandit task we estimated reward rate using a time-based leaky-integrator. Reward rate was incremented by 1 at each time the rat received a reward, and exponentially decayed with time constant τ. τ was varied between 1-2500s, to find the strongest negative correlation between reward rate and the DA peak after Side-In (within 0-1s, on rewarded trials; i.e. positive RPE coding). To estimate learning rate, we used a trial-based delta-rule. This model tracks a state value that is updated once per trial by $V(t) = V(t-1) + \alpha * (r - V(t-1))$; V(t) is the trial-based state value at trial t, α is the learning rate and r is the outcome of each trial (0 or 1). By varying the value of α between 0 and 1 (in 0.01 steps) we found an optimal value for each DA signal that would minimize the correlation between state value and peak DA signal in a 1s window after Side-in.

Real-time models. The CSC model is a standard temporal-difference model of conditioning (49). Values are defined as a linear function of features x and weights w, $V_t(x) = w_t x = \sum_{i=1}^{n} w_t(i)x(i)$, where n is the time steps in a trial. The vector x is non-zero only at the t_{th} element at time step t after cue onset, i.e., $x(i) = \delta_{it}$, where δ_{it} is the Kronecker delta function. In addition to activating a single distinct feature for each cue, we also included shared features activated by any of the three cues, to allow for generalization. In the results presented we used a single shared feature, but increasing the number of shared features did not qualitatively affect results (not shown). The weights w update according to $w_{(t+1)} = w_t + \alpha\delta_t e_t$, where α is the learning rate (we used $\alpha = 0.01$) , δ_t is the RPE and e_t is an eligibility trace. The RPE is defined as $\delta_t = r_t + \gamma V_t(x_t) - V_t(x_{(t-1)})$, where γ is the discounting factor. The eligibility trace e_t is included to accelerate learning and updated by $e_{t+1} = \gamma\lambda e_t + x_t$, where λ is a decay factor (we used $\lambda = 0.98$). The CSC model was run separately for each discount factor.

The RNN model, based on an advantage actor-critic architecture (87), is composed of LSTM units (88). These are organized as three sub-networks ("DLS", "DMS", "VS") of 32 nodes each, with internal recurrent connections but without direct connections between sub-networks. Each sub-network receives the same copy of the sensory inputs at each time point, and generates its own value estimate using a distinct discount factor. All three sub-networks project to the same policy component, together generating the probability for taking an action (either "poke" or "no-poke"). These probabilities are sampled to determine the action at each time step. We used a time step of 100 ms.

The sensory inputs include the food delivery click (one input with 0 for no-click and 5 for click), auditory cues, and background dimensions. Background dimensions (3 in number) are included to mimic the background or contextual inputs to the network and are all set constantly to 1. The auditory cues consist of 20 inputs, of which 3 inputs are the distinctive one-hot features of the cues and the remainder are set to 1 during all cue presentations to produce similarity between cues.

At each time step the RNN model receives reward feedback. Before reward delivery, the reward is 0 for taking the action "non-poke", and -0.006 for taking the action "poke", i.e., there is a small poking cost to discourage constant poking. If the poke output is maintained on consecutive time steps, the cost is reduced to 10% of that for first poke. After the reward delivery click in a rewarded trial, the reward is presented with a delay of 3 steps and the reward received for the first poke after the delay is 1.0.

The network was trained to perform the conditioning task by minimizing a loss function with three terms,

$$L^{\theta}_{PPO} = E_t[L^P_t(\theta) + \beta_V L^V_t(\theta) - \beta_e L^e_t(\theta)]$$

where the expectation was over a sequence of time steps with length T. We used T = 5000 steps, which encompasses multiple (~ 20) trials. We took the proximal policy optimization (PPO) for estimating the policy loss, which has the following form (89)

$$L^p_t(\theta) = min(\rho_t A_t, clip(\rho_t, 1-\epsilon, 1+\epsilon)A_t)$$

where $\rho_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$is the probability ratio, whose value is clipped with a parameter ϵ. The advantage A_t includes three components,

$$A_t = A^{GAE}_{VS}(t) + A^{GAE}_{DMS}(t) + A^{GAE}_{DLS}(t)$$

where each term is the generalized advantage estimator (GAE) (90) from one of the three sub-networks. Take the VS term as an example and define $\delta^{VS}_t = r_t + \gamma_{VS}V^{VS}_{t+1} - V^{VS}_t$ as the RPE at time t, then

$$A^{GAE}_{VS}(t) = \delta_t + (\gamma_{VS}\lambda)\delta_{t+1} + \ldots + (\gamma_{VS}\lambda)^{T-t}\delta_T$$

where T is the sequence length and λ is a parameter for GAE.

The value loss was given by

$$L^V_t = (\overline{r}^{VS}_t - V^{VS}_t(\theta))^2 + (\overline{r}^{DMS}_t - V^{DMS}_t(\theta))^2 + (\overline{r}^{DLS}_t - V^{DLS}_t(\theta))^2$$

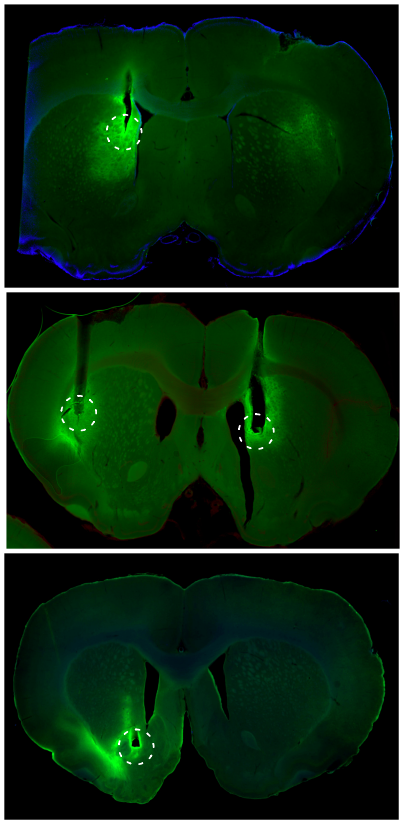
where $\overline{r}^{VS}_t$, $\overline{r}^{DMS}_t$, $\overline{r}^{DLS}_t$ are the expected discounted rewards within the sequence, given the corresponding discount factor for each subnetwork. We used the value right after T to bootstrap the contribution from rewards beyond this sequence. For instance, the expected reward for VS has the following expression

$$\overline{r}^{VS}_t = r_t + \gamma_{VS}r_{t+1} + \ldots + \gamma^{T-1}_{VS}r_{t+T-1} + \gamma^T_{VS}V_{t+T}$$

Since γ_{VS} is very close to 1, the expected reward for "VS" sub-network reflects contributions from multiple trials. Faster discounting for "DMS" and (especially) "DLS" sub-networks results in minimal contributions from subsequent trials. The entropy term L^e represents the entropy of the probability distribution of taking the two actions and was added to encourage the exploration. The parameters used were: $\beta_V = 0.8$, $\beta_e = 0.001$, $\gamma_{VS} = 0.9999$, $\gamma_{DMS} = 0.99$, $\gamma_{DLS} = 0.95$, $\lambda = 0.98$. The weights of the network were updated using Adam method (91), with learning rate 0.0005.

Contributions. A.M. performed the behavioral photometry experiments and bandit task analyses. W.W. performed the computational modeling and Pavlovian task analyses. J.B. developed the conceptual framework, oversaw the study, and wrote the manuscript.

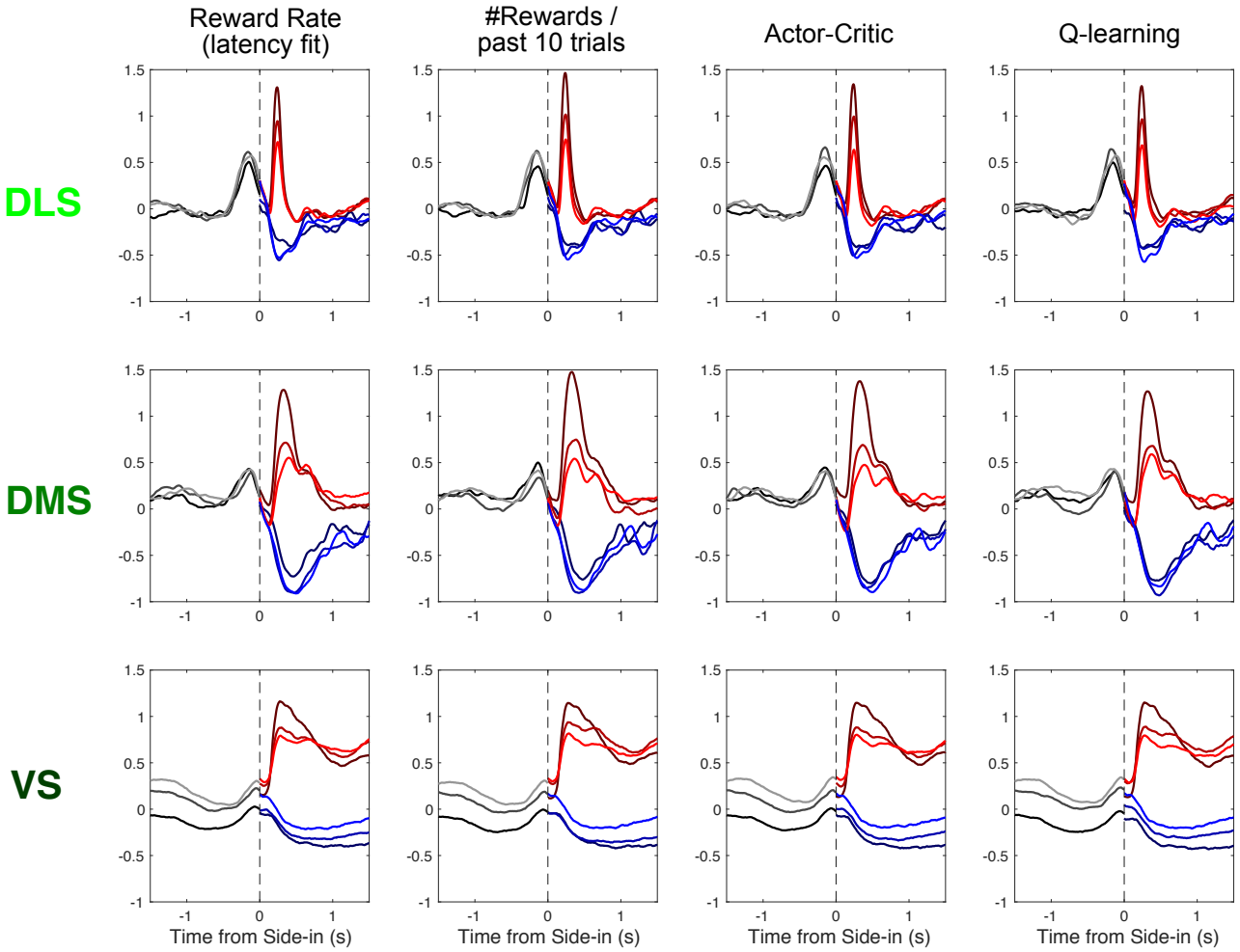
A



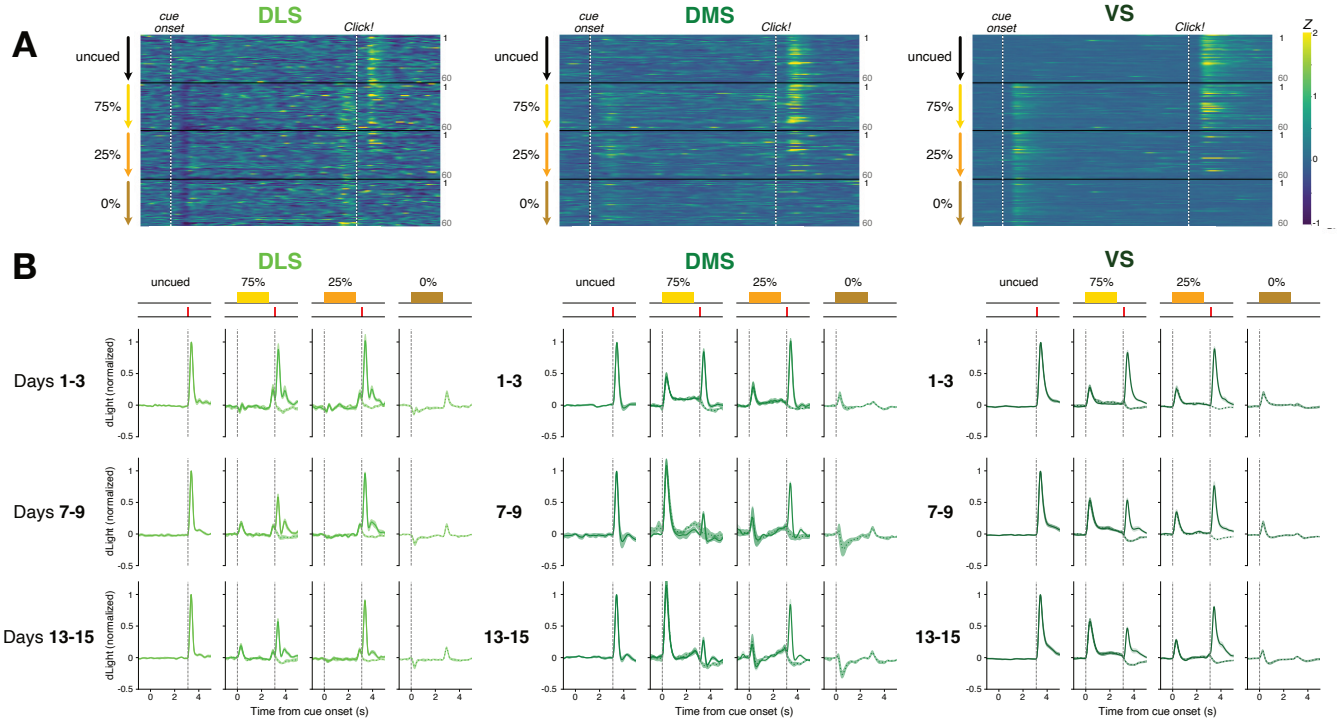
B

	Bandit			Pavlovian		
Rat #	DLS	DMS	VS	DLS	DMS	VS
1065			*			
1066			**			
1088			*			
1089			*			
1105			*			
1106			**			
1107			*			
1277	*	*		*	*	
1278	*	*		*	*	
1299	*	*		*		
1300	*	*		*	*	
1301	*	*		*		
1358						**
1359						**
1366						*
1381						*
1382					*	*

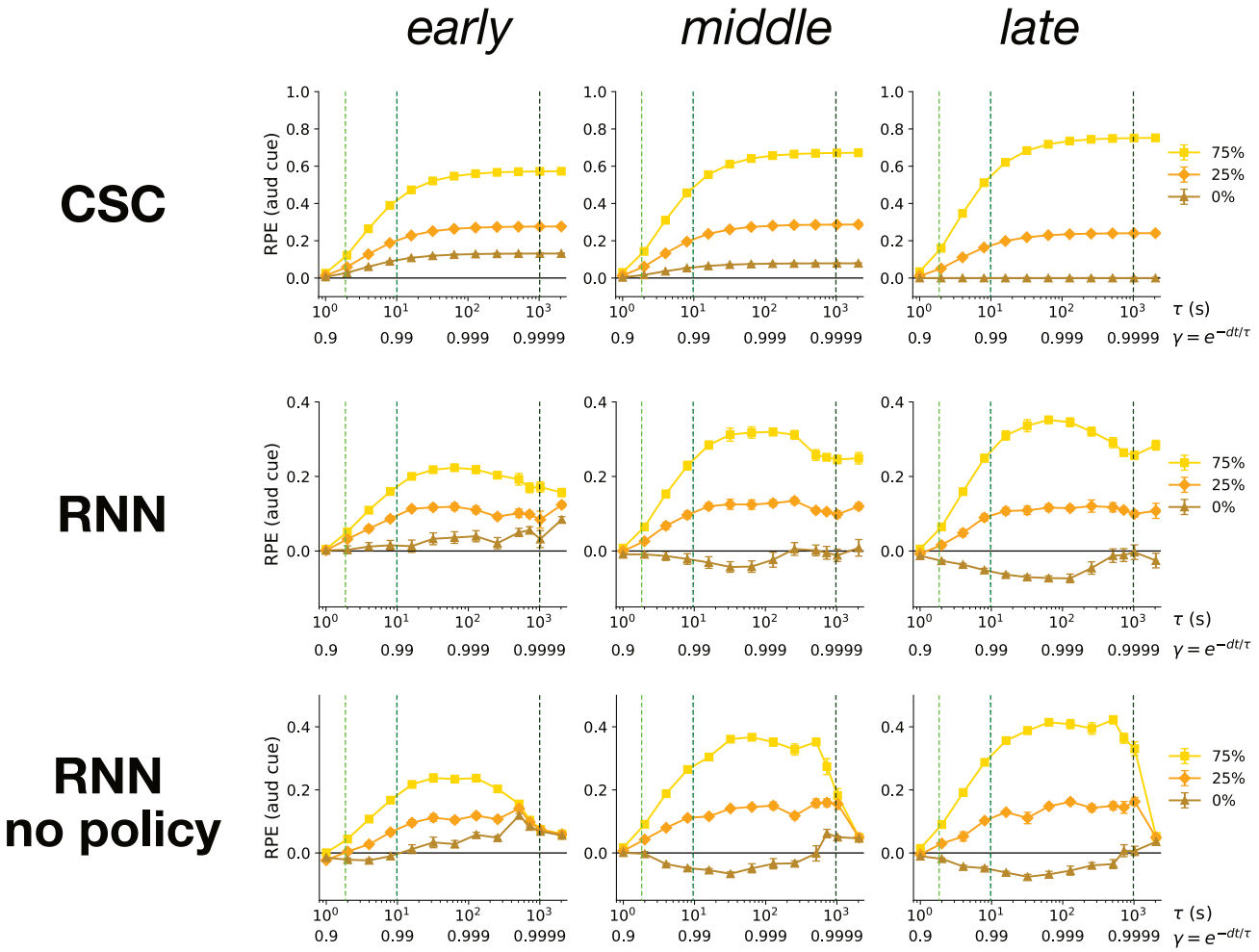
Supplementary Figure 1. Photometry recording locations. **A**, Histology examples showing optic fiber tip locations (circled) and dLight1.3b expression (green), in DLS (top), DMS (middle), VS (bottom). **B**, Table showing distribution of fiber subregions included for each task. Double asterisk indicates bilateral recording from the same subregion. Data from rats 1065-1107 were previously reported(12)



Supplementary Figure 2. RPE patterns with alternative value estimators. Each column shows the same data and format as Fig. 2B, but broken down using different ways of estimating rats' reward expectation. From left, "Reward Rate" uses a leaky integrator as in Fig.2, but this time choosing time constant τ to produce the strongest (negative) correlation between reward rate and behavioral latency to initiate the trial (as in (12, 27). "Rewards in the past 10 trials" is a simple count. "Actor-Critic" uses the Critic value from a trial-based Actor-Critic model, fitting the Critic learning rate to behavioral latency and the Actor α, β parameters to left and right choices. Q-learning uses a trial-based Q-value model, fitting the α and β parameters to choices and using Q (chosen action) as reward expectation.



Supplementary Figure 3. Distinct development of DA cue responses in each subregion. A, single-animal examples showing DA signals on each trial in the first Pavlovian session, including the very first exposure to the 75, 25, 0% cues. Note that in the DLS example the response to all the novel cues is negative, while in the VS example all responses are positive. **B**, Average responses in each subregion at three different learning stages (days 1-3, 7-9 or 13-15). In all subregions discrimination between cues increases with time, but this is slow in VS.



Supplementary Figure 4. Effects of extended model training on cue discrimination with different discount factors. Top row, cue-evoked RPEs in the CSC model at "early" (600 training steps), "middle" (1000) and "late" (3800) stages of learning, as a function of γ, or equivalently the time parameter τ. ($\gamma = e^{-dt/\tau}$, where dt is the time step size, here 100ms). Green dashed lines mark $\gamma = 0.95$, 0.99, and 0.9999. Note that for low γ all cue responses are small even after learning, since any potential reward is heavily discounted. This CSC model initially shows a positive response to the 0% cue due to overlapping cue representations; over training this response fades to zero (but cannot become negative). Middle row, same for an RNN model (early = 100, middle = 500, late = 900 training steps). To isolate the effect of varying γ, this model variant used just a single network (a single γ) rather than three. Note that at early and middle stages of learning, if γ is close to 1 the RNN model shows less discrimination between cues compared to intermediate γ, consistent with the observed difference between VS and DMS. Bottom row, same as middle row, but also removing the Actor (poking) component.