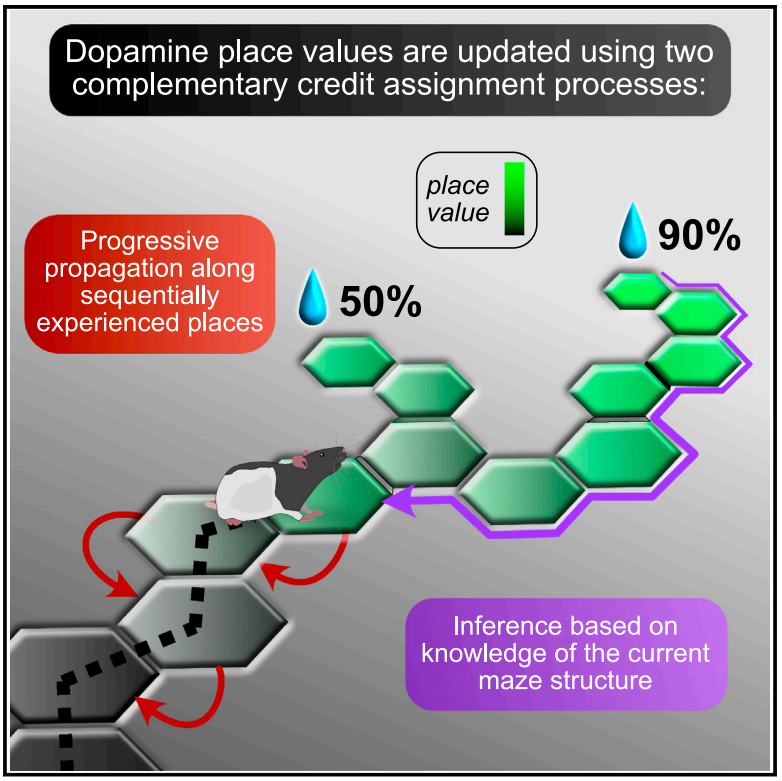
Neuron

Article

Dual credit assignment processes underlie dopamine signals in a complex spatial environment

Graphical abstract



Highlights

- Accumbens dopamine scales with the evolving values of maze locations
- Dopamine place values propagate between sequentially experienced locations
- Maze knowledge allows the inference of updated values even without direct experience
- The discovery of novel path opportunities elicits dopamine transients

Authors

Timothy A. Krausz, Alison E. Comrie, Ari E. Kahn, Loren M. Frank, Nathaniel D. Daw, Joshua D. Berke

Correspondence

joshua.berke@ucsf.edu

In brief

Krausz et al. investigate how expectations of future reward are updated through experience. In rats traversing a complex maze, they show that nucleus accumbens dopamine scales with reward expectation from each location. This expectation signal propagates between adjacent spatial locations and is also inferred using knowledge of maze structure.

Krausz et al., 2023, Neuron *111*, 3465–3478
November 1, 2023
https://doi.org/10.1016/j.neuron.2023.07.017

CellPress

Neuron

CellPress
OPEN ACCESS

Article

Dual credit assignment processes underlie dopamine signals in a complex spatial environment

Timothy A. Krausz,[1] Alison E. Comrie,[1] Ari E. Kahn,[4] Loren M. Frank,[1,3,5] Nathaniel D. Daw,[4] and Joshua D. Berke[1,2,6,7,*]
[1]Neuroscience Graduate Program, University of California, San Francisco, San Francisco, CA 94158, USA
[2]Kavli Institute for Fundamental Neuroscience, and Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA
[3]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[4]Department of Psychology, and Princeton Neuroscience Institute, Princeton University, Princeton, Princeton, NJ 08544, USA
[5]Department of Physiology, University of California, San Francisco, San Francisco, CA 94158, USA
[6]Department of Neurology and Department of Psychiatry and Behavioral Science, University of California, San Francisco, San Francisco, CA 94158, USA
[7]Lead contact
*Correspondence: joshua.berke@ucsf.edu
https://doi.org/10.1016/j.neuron.2023.07.017

SUMMARY

Animals frequently make decisions based on expectations of future reward ("values"). Values are updated by ongoing experience: places and choices that result in reward are assigned greater value. Yet, the specific algorithms used by the brain for such credit assignment remain unclear. We monitored accumbens dopamine as rats foraged for rewards in a complex, changing environment. We observed brief dopamine pulses both at reward receipt (scaling with prediction error) and at novel path opportunities. Dopamine also ramped up as rats ran toward reward ports, in proportion to the value at each location. By examining the evolution of these dopamine place-value signals, we found evidence for two distinct update processes: progressive propagation of value along taken paths, as in temporal difference learning, and inference of value throughout the maze, using internal models. Our results demonstrate that within rich, naturalistic environments dopamine conveys place values that are updated via multiple, complementary learning algorithms.

INTRODUCTION

Animals frequently make motivated choices based on prior experiences—for example, selecting paths toward locations where food was previously found. Achieving such adaptive decision-making can pose a computational challenge. In particular, decision points can be separated from rewards by considerable gaps in time and space. When rewards are obtained (or unexpectedly omitted), this separation produces a "credit assignment problem": determining which places and choices ought to gain or lose value. The specific algorithms that brains use to solve this problem are not well understood.

Reinforcement learning (RL) theory provides an array of candidate algorithms for generating and updating value signals.[1] In "temporal difference" (TD) learning, value is passed between sequentially experienced states (situations). In brief, as each state is encountered its associated value becomes eligible for updating. Unexpected rewards, or transitions to states with unexpected values, evoke "reward prediction errors" (RPEs). RPEs are learning signals: they update the values of eligible states. In this way, values can be progressively propagated back to earlier states, over repeated episodes of experience. TD RPEs can be encoded by brief (phasic) changes in the firing of midbrain dopamine (DA) cells[2–5] and by corresponding changes in DA release in the nucleus accumbens (NAc).[5,6] However, despite the compelling correspondence between phasic DA and TD RPEs, current evidence that value propagates along sequences of states in a TD-like manner is sparse at best.[7–9]

TD learning is a "model-free" (MF) algorithm: learning occurs only from direct experience of states, without using knowledge of how those states are related. A complementary set of "model-based" (MB) algorithms can achieve greater flexibility in learning and decision-making, by using knowledge about state relationships to infer and update values. For example, after taking one path and receiving reward, MB algorithms can increase values along *alternative* paths to the same reward location.[10,11] In at least some behavioral contexts, DA signals reflect RPEs that incorporate such inferred information.[12–15]

NAc DA release also ramps up as animals actively approach expected rewards.[5,16–19] These ramps appear to signal the value of the upcoming reward, discounted by spatial distance (although they have also been interpreted as RPEs[20,21]). As DA ramps are more apparent when the behavioral context favors

use of an internal model,[22] they have been proposed to reflect ongoing MB calculations.

Yet overall, existing evidence does not tease apart the specific algorithms used to estimate and update values, nor does it reveal how these values are reflected in DA signals. Many behavioral tasks commonly used to investigate DA and value coding involve only minimal separation between an action and its outcome (e.g., Hamid et al.,[17] Morris et al.,[23] and Parker et al.[24]), thus avoiding the challenging credit assignment question. In other paradigms, applying RL ideas involves unsupported arbitrary assumptions[25]—e.g., choosing the set of discrete covert states to span a time interval between a cue and reward.[2,9] Spatial tasks have the advantage that the brain has a well-studied set of spatial representations that could serve as a basis for RL states.[26] However, most spatial tasks—especially those in which DA dynamics have been investigated—are very simple (e.g., T-mazes[19,27]). This simplicity is often useful, but it can prevent critical tests that distinguish between credit assignment algorithms.

To better elucidate neural credit assignment processes within natural environments, we developed a dynamic, complex spatial foraging task for rats. In this task, animals traverse through numerous distinct decision points in the pursuit of reward, and choices are separated from their outcomes by multiple steps in space and time. Furthermore, reward contingencies can be unstable, and the available paths to reward locations can be unexpectedly reconfigured. We show that rats readily adapt to these changes and incorporate both costs (current distances to reward ports) and benefits (current reward probabilities) into their decisions.

Using fiber photometry, we observe DA RPE coding at reward receipt and also strong DA pulses when rats discover newly available paths. We confirm that NAc DA ramps up with reward approach and show that these ramps reflect a robust relationship between DA release at each location and the dynamically changing value of that location. We then take advantage of this DA place-value signal to examine how values are updated from trial to trial. We report strong evidence for *both* MF TD-like local propagation of values between adjacent locations and MB global inference of values throughout the maze environment.

RESULTS

Cost-benefit decision-making in a novel maze task

The maze (Figure 1; Video S1) is triangular with a reward port at each corner, each with a distinct reward probability.[17,28–31] The available paths to these reward ports are defined by a set of barriers, constraining rats into making sequences of left and right turns from each "hex" location. The task is self-paced, that is, the end location for each "trial" is the start for the next, and each reward port can be approached from multiple starting locations. Overall, rats (n = 10) were more likely to choose a port if it had a higher probability of reward (Figure 1B) and was closer (Figure 1C), compared with the alternative. A mixed-effects logistic multiple regression, incorporating any turn biases (see STAR Methods), revealed highly significant effects of both reward probability (mean $\beta = 1.605 \pm 0.163$ SEM, $p = 5.31 \times 10^{-23}$) and distance cost (mean $\beta = -6.805 \pm 0.550$ SEM, $p = 3.46 \times 10^{-35}$) on port choices (Figure 1D). After each block of 50–70 trials (traversals between ports), either the reward probabilities changed (Figure 1E) or a barrier was moved to change available paths (Figure 1F). After a change in reward probabilities, rats increased their choice of ports whose reward probability had increased (Figure 1G). Following a barrier move, rats adjusted their port choices to favor shorter paths (Figure 1H) and also progressively refined their specific paths to be more efficient (Figure S1).

Phasic dopamine responses to rewards and novel path opportunities

During task performance we recorded NAc DA dynamics using fiber photometry with the fluorescent DA sensor, dLight1.3b[32] (n = 10 rats, 19 fiber locations, 82 behavioral sessions, 296 blocks, 16,379 trials, mean of 1,638 trials per rat). We first examined DA changes around reward port entry, since receipt (and omission) of probabilistic reward is an obvious time to look for the best-known correlate of NAc DA, RPE signals. DA transiently increased or decreased depending on whether reward was delivered or omitted, respectively (Figure 2B). The magnitude of these phasic changes depended on port reward probability, in a direction consistent with RPE coding (Figure 2C, Pearson correlation, rewarded trials mean coefficient = -0.221 ± 0.098 SD; omission trials mean coefficient = -0.111 ± 0.062 SD; both significantly different to zero across n = 10 rats, two-tailed Wilcoxon signed rank tests, $p = 1.95 \times 10^{-3}$ each). To better estimate RPE at the single-trial level, we fit a simple trial-level RL algorithm to rats' port choices and reward outcomes ("Q learning"; see STAR Methods). DA following port entry significantly scaled with these RPE estimates (Figure S2), although encoding of positive RPEs was notably stronger and more consistent across rats, compared with negative RPEs (in line with prior studies[3,5,33]).

We also observed large phasic increases in DA when rats first encountered a newly available hex—i.e., where a barrier had been previously located, but not anymore (Figures 2D–2F). This was not simply a response to any unexpected sensory event, since encountering a newly *blocked* hex resulted in a significantly smaller or absent DA pulse (Figure 2F). Additional analyses suggested that the response to newly available hexes is larger on trials in which rats chose to take the new path, rather than ignoring it (Figure S2C), and when the newly available hex was closer to the final destination port (Figure S2D). However, these latter observations would require a larger dataset of new path discoveries for solid statistical support.

Dopamine ramps reflect expectations of upcoming reward

We next examined whether the reward approach ramps previously reported for NAc DA are also present in this more complex spatial environment. Average NAc DA indeed ramped up within each trial, until shortly before arrival at the reward port (Figure 3A). This overall ramp was significantly positive in 9 of 10 individual animals (16/19 individual fibers; Figure S3A). To better understand the computations that give rise to this ramp, subsequent analyses focused on those nine rats. The magnitude of the

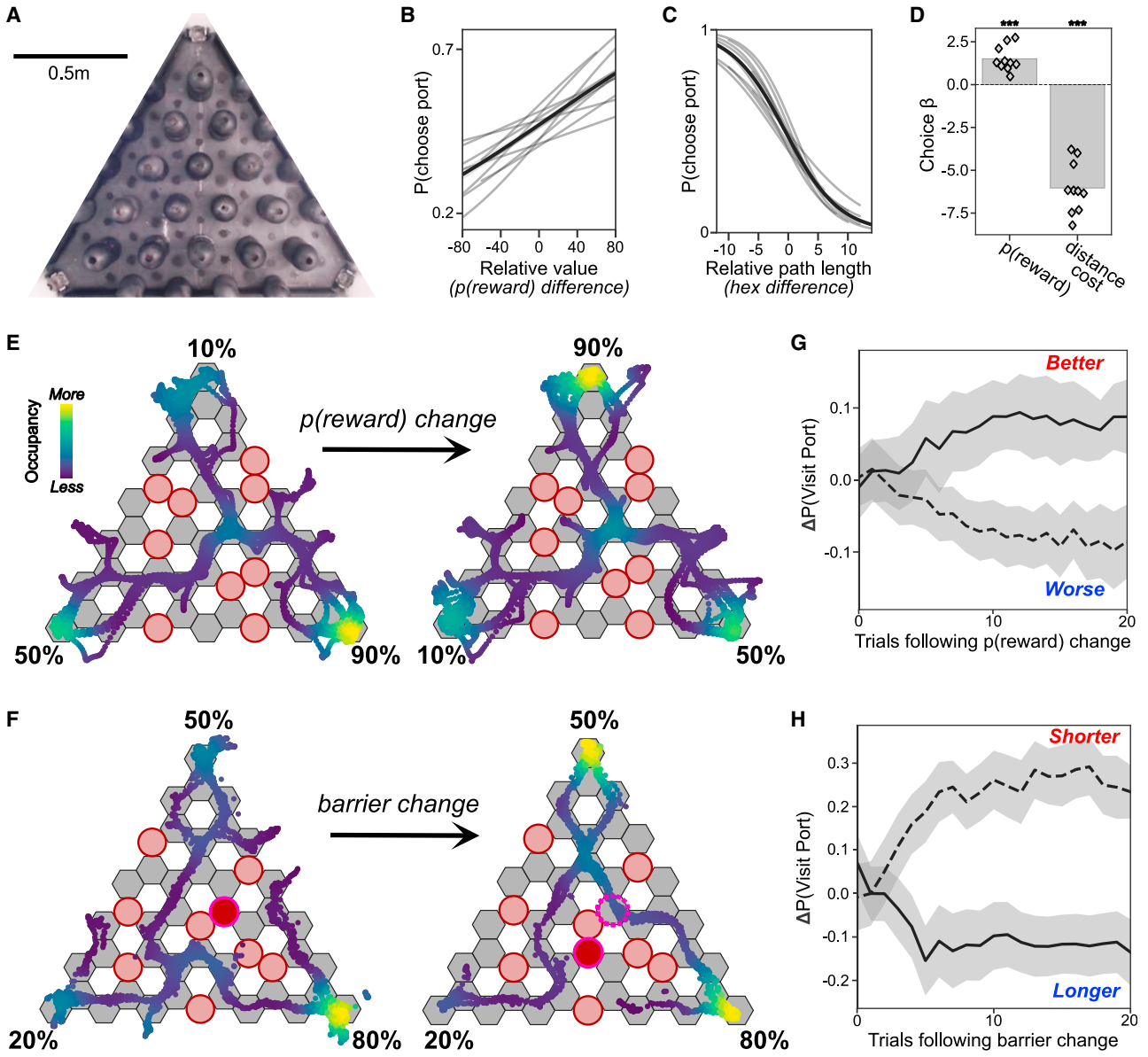


Figure 1. Adaptive behavior in the spatial foraging task

(A) Bird's-eye view of the maze. Permanent barriers (black columns) divide the area into 49 hexagon-shaped choice points ("hexes"). Additional movable barriers (absent here) determine the available paths to the reward ports at each corner. Once visited, a port's reward probability becomes zero until another port is visited.

(B) Probability of choosing an available port as a function of the difference between that port's and the alternative port's reward probabilities. Gray traces are individual rat logistic curves fit to the data, and the black line shows the mean relationship.

(C) Same as (B), but a function of the difference between path lengths to the available ports.

(D) Results of logistic multiple regressions run for each individual rat, showing the positive influence of reward probability and the negative influence of path length on choices. Significance asterisks are from the mixed-effects regression analysis. For (B)–(D), only the second half of each block (trial number > 25) was included to allow rats time to adapt to changes (n = 10 rats, 82 sessions, 9,079 trials).

(E) Example of a reward probability change. Red circles indicate hexes containing a movable barrier; dots show the rat's detected positions (color coded by occupation density; second halves of blocks). Empty white hexes indicate the positions of the permanent barriers shown in (A).

(F) Example of a barrier change. The dark red circle with a pink outline shows the moved barrier.

(G) Mean change in port choice probability following increases (solid line) or decreases (dashed line) in reward probability (n = 10 rats, 36 sessions, 134 blocks; error bands indicate ±SEM).

(H) Mean change in port choice probability following increases (solid line) or decreases (dashed line) in the path length to reach the goal (n = the same 10 rats, 46 sessions, 162 blocks; error bands indicate ±SEM). "Trials" in (G) and (H) include only those where the rat had the opportunity to choose the port in question.

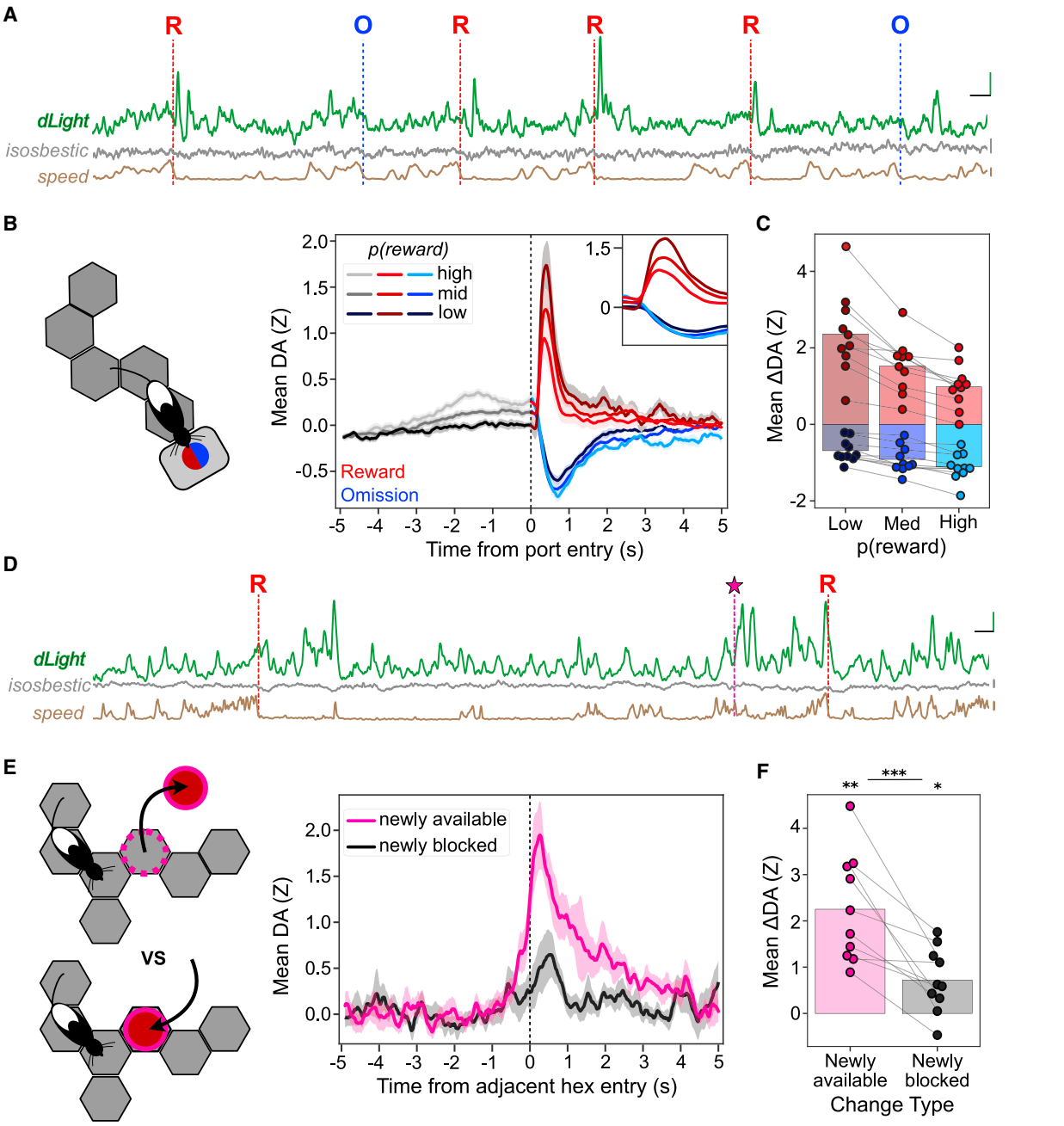


Figure 2. Dopamine pulses at rewards and novel path opportunities

(A) Example trace of dLight, isosbestic (405 nm) control signal, and running speed over three trials. Red "R"s indicate reward deliveries, and blue "O"s indicate reward omissions upon port entry. Vertical scale bars indicate 2 *Z* for fluorescence signals and 20 cm/s for speed. The horizontal scale bar indicates 2 s of time.

(B) Left, cartoon of rat arriving at port. Right, average DA (*Z* scored) aligned to the port entry, pooled by the destination port's reward probability ("high" = 80% or 90%; "medium" = 50%; "low" = 10% or 20%). Traces are separated into rewarded (red) or omissions (blue) following port entry, and error bands indicate ±SEM (n = 10 rats). Inset shows close up of the first 1 s after port entry. Only the second half of each block (trial number > 25) was included (82 sessions, 9,079 trials).

(C) Mean change in DA as a function of port reward probability, separated by rewarded (red) and unrewarded (blue) trials. Changes in DA measured as peak DA within 0.5 s following reward and minimum DA within 1 s following omission, subtracting instantaneous DA at port entry.

(D) Example trace of dLight and running speed across three trials, including when the rat discovered a newly available path (pink star). Scale bars as in (A).

(E) Left, cartoon of rat discovering the absence (top) or presence (bottom) of a barrier. Right, mean DA on each of these trial types; error bands indicate ±SEM. DA signal is aligned on entry into the hex adjacent to the newly changed hex (pink, newly available; black, newly blocked; each n = 10 rats, 106 events).

(legend continued on next page)

DA ramp scaled with the current reward probability of the approached port (Figure 3B), consistent with DA tracking the rats' evolving expectations of receiving reward on the current trial. We therefore assessed how DA ramping during port approach is affected by whether that port was rewarded or not at the last visit (Figure 3C). DA ramps were stronger when the destination port had been most recently rewarded and weaker following an omission. This effect was significant along the full length of the path (note asterisks in Figure 3C), not just the hexes closest to reward. To rule out non-specific effects of recent rewards on DA signals, we performed a multiple regression analysis comparing the impact of the most recent reward outcome at each of the three ports (Figure 3D). DA ramps selectively reflected reward history for the port at the end of the path taken on the current trial, rather than (for example) tracking overall recent reward rate,[34] or the history of rewards for both potential destination ports together. Average running speed was also greater as rats ran toward higher-probability ports (Figure 3B). However, running speed peaked later than DA (Figures 3A and 3B), and cross-correlograms suggested that DA was predictive of speed (potentially driving the vigor of running) rather than merely reflecting it (Figure S3B).

A spatial map of value

These ramping dynamics during port approach suggest that DA at each maze location may signal the rats' evolving expectation of receiving reward, discounted by spatial distance. To assess this "place-value" possibility, we turned to models that generate reward expectation estimates for each specific spatial location (at entry into each hex, from each direction; 126 distinct hex states). As a first pass, we again applied a simple learning algorithm that tracks experienced reward probabilities at each port (Figure 3E), but we then distributed these values, discounted by spatial distance, throughout the maze ("value iteration";[1,35] Figure 3F; see STAR Methods). The resulting hex-level pattern of value closely resembled DA on each trial (Figure 3G), and a mixed-effects multiple regression analysis revealed a highly significant relationship between DA and these hex values (Figure 3H, $p < 0.0001$, likelihood ratio test, chi-squared distribution with 77 degrees of freedom to account for each session-optimized γ value; see STAR Methods). This regression analysis also included running speed, yet hex values accounted for much more of the explained variability in the DA signal (Figure 3I).

Over repeated trials, DA signals propagate backward along taken paths

This value map provides a reasonable first approximation to DA signals as rats run through the maze. However, the value-iteration algorithm requires perfect knowledge of current maze structure, together with the immediate and complete distribution of value updates to all hex states on every trial. Rat brains might actually use less computationally demanding algorithms to generate place values. These algorithms could produce telltale signatures in value coding while foraging—including deviations from smooth ramps.

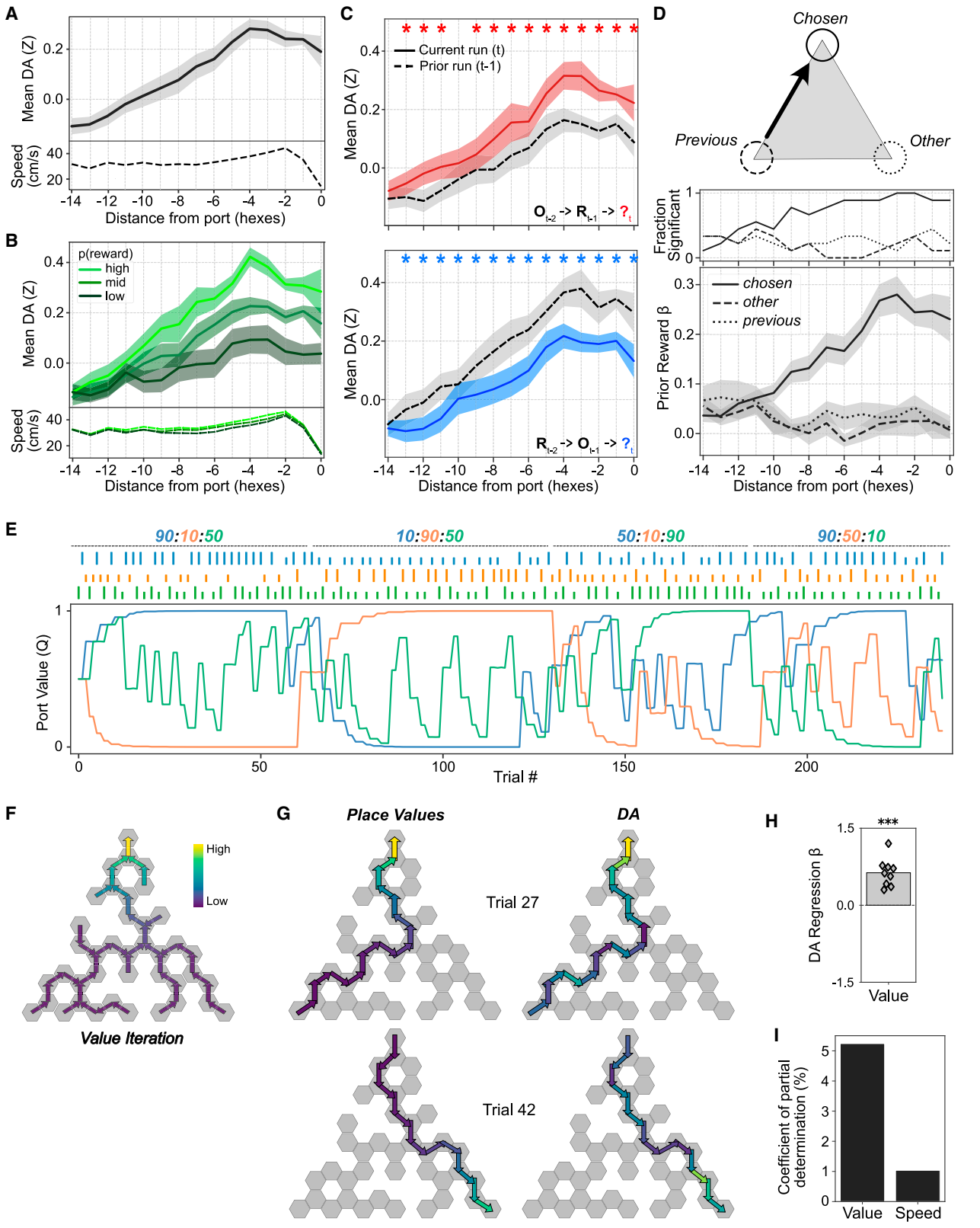
First, we looked for evidence of TD learning, as this has been an especially prominent framework for interpreting DA signals in simpler settings. In its most basic form, TD(0) (also called "one-step" TD), RPEs update only the values associated with the immediately preceding state[1] (Figure 4A). Therefore, when a sequence of states results in an unexpected reward, earlier states in the sequence do not receive value updates right away. Instead, updates progressively propagate backward along the sequence, over multiple episodes of experience. This type of learning rule has a clear signature: values of states more distant from reward should depend on reward outcomes in the more distant past, rather than the most recent outcomes.

TD can also propagate value more rapidly by maintaining memory traces for recently visited locations and using these to determine eligibility for later value updates. Such an algorithm is referred to as TD(λ).[1,7,35] By altering the eligibility trace decay parameter λ, value updates can be restricted to the single preceding state ($\lambda = 0$, as above) or, at the other extreme, cover the entirety of the experienced path ($\lambda = 1$).

The resulting difference in value dynamics can be clearly illustrated by considering the impact of a single reward, among a series of omission trials for the same path (Figure 4B). In simulations (see STAR Methods), with TD(0) the reward evokes a value bump that propagates backward over the course of multiple traversals (Figure 4C). By contrast, with TD(1) value is immediately updated across the full traversed path, so that outcomes simply change the gain of the ramping value function (Figure 4D). To broaden this analysis to include all sequences of reward outcomes, we turned to multiple regression. We examined how values at each location along a path depend upon prior reward outcomes. Specifically, we performed a multiple regression of how the path's prior five reward outcomes affect value at each distance from the reward port (Figures 4E and 4F). We then identified the place along a path where each prior reward had its maximum effect (regression coefficient) on place value. As expected, in a TD(0) simulation, the information from older reward outcomes had its strongest influence on value farther away from the reward port (Figure 4E), in stark contrast to TD(1) (Figure 4F).

We then applied the same analyses to DA signals. First, we examined DA dynamics after rats experienced one reward among a series of omissions for traversing the same path (as in Figure 4B). The reward appeared to cause a spatial bump in DA, which moved further back from the reward port over successive traversals (Figure 4G)—i.e., the key signature of TD learning with low λ. Next, we performed the multiple regression with all trial sequences, as in Figures 4E and 4F, but with observed DA signals. This analysis resulted in a pattern resembling TD(0): older outcomes had the largest influence on DA signals farther from the reward location (Figure 4H; two-tailed Wilcoxon signed rank, $p = 1.95 \times 10^{-2}$). This provides clear evidence that updates of DA value signals incorporate TD(0)-like progressive, backward propagation.

(F) Mean change in DA (peak DA within 1 s following novel hex discovery—DA 1 s before novel hex discovery) was significantly higher for newly available than for newly blocked hexes (available vs. blocked: $p = 9.76 \times 10^{-4}$, one-tailed paired Wilcoxon signed rank test; available vs. 0: $p = 1.95 \times 10^{-3}$, two-tailed Wilcoxon signed rank test; blocked vs. 0: $p = 0.014$, two-tailed Wilcoxon signed rank test). Individual rat means are plotted as dots. Bars represent means over rats. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.



(legend on next page)

No single algorithm is likely to explain both this evidence for value propagation and the path-wide shifts in DA ramps following reward or omission described earlier (Figure 3C). Although each of these can arise separately as the extreme cases of TD(λ) (i.e., λ close to 0 or 1, respectively), there is no intermediate setting of λ at which both these patterns co-occur. Consistent with this, fitting a TD(λ) hex-state RL algorithm (see STAR Methods) to the observed DA data could model the broad shifts, but the resulting large λ failed to also reproduce the progressive propagation of DA and its dependence on reward history (Figures S4A–S4E). Fit λ numbers were consistently high for individual sessions (Figure S4B), ruling out the possibility that our results reflect variability in λ across time or between animals.

We therefore explored the possibility that multiple credit assignment algorithms, operating over distinct spatial scales, could collectively update DA value signals. To this end, we first built a model that learns through a mixture of TD(0) and TD(1). As expected, value ramps in this combined model superimposed bumps and broad shifts (Figure 4I). This combined model also shows the same pattern as DA and TD(0) in regression analysis, namely, the increasing distance of maximum impact of rewards earlier in time (Figure 4J).

We reasoned that progressive propagation of DA values should be more apparent if we were to remove the broad shifts in the DA signal. We did this by modeling each trial's ramp as a linear scaling of the average DA ramp (see STAR Methods). As expected, removing the overall ramp left a residual DA signal that propagated backward along the path over trials (Figures 4K and 4L). These results are consistent with updating of DA value signals by at least two mechanisms—a TD(0)-like process responsible for backward signal propagation and a second process capable of shifting the whole ramp at once.

DA place values are also globally updated through inference

Furthermore, the behavioral choices of the rats were more sophisticated than would be expected from MF TD alone. In the maze, each reward port can be reached from multiple starting points (Figure 5A). MF TD learning would only update values along the path that was actually taken. However, we found that reward at a given port increased rats' likelihood of choosing that same port at the next opportunity, *both* when the rat previously took the same path ($p = 9.77 \times 10^{-4}$, two-tailed Wilcoxon signed rank test) or an alternative path ($p = 4.88 \times 10^{-3}$, two-tailed Wilcoxon signed rank test) to that port (Figure 5B). A potential confound could arise from correlations between this most recent reward outcome and prior reward outcomes at that same port, for which the rat may have taken the same path. To control for this, as well as any turn-direction bias, we conducted a mixed-effects multiple regression analysis and included the past five reward outcomes as features (see STAR Methods). We confirmed that a previous reward at a port made current choice of that port more likely, both when the rat had taken the same path to obtain reward ($p = 2.26 \times 10^{-6}$) or an alternative ($p = 0.0242$). This suggests the use of MB algorithms to infer that hexes along alternative paths to that same reward location have also changed value.

We therefore assessed whether DA ramps similarly rely upon MB processing and knowledge of maze structure. We confined this analysis to the critical "path-dependent" hexes—those that have no overlap with other paths to the same port (Figure 5A). We found that a prior reward at a port results in elevated DA in these path-dependent hexes, both when the rat previously took the same path ($p = 3.90 \times 10^{-3}$, two-tailed Wilcoxon signed rank test) or an alternative path ($p = 3.90 \times 10^{-3}$, two-tailed Wilcoxon signed rank test) to that port (Figure 5C). Once again, to control for the possibility that this result reflects experiences on even earlier trials, we ran a regression analysis and included the prior five reward outcomes (see STAR Methods). DA still displayed a significant relationship with the most recent reward outcome at the goal port, both when the rat previously took the same path ($p = 5.80 \times 10^{-3}$) or an alternative path ($p = 0.0190$) to the goal port. Consistent with this, the effect of prior reward (or omission) on DA ramping was observed whether

Figure 3. DA ramps reflect dynamic expectations of upcoming reward

(A) Mean hex-level running speed (bottom) and DA (top) as rats approached the reward ports (n = 10 rats, 82 sessions, 15,918 trials), as a function of distance.

(B) Mean hex-level running speed (bottom) and DA (top) during port approach, pooled by p(reward) of the destination port. Only the second half (trial number > 25) of each block was included (n = 9 rats, 70 sessions, 7,614 trials).

(C) Examining the effects of reward on DA ramping along successive runs to the same port. Dashed lines indicate the prior run to the port (t−1), and solid lines indicate the current run to the port (t). Top, mean DA over successive runs to the same port, where reward was omitted two visits ago (t−2), but reward was delivered on the prior visit (t−1; n = 9 rats, 1,935 sequences). Red asterisks indicate a significant increase in hex-level DA ($p < 0.05$, one-tailed Wilcoxon signed rank test). "R" and "O" denote rewards and omissions, respectively, on the t−n previous visits to the port. Bottom, same as top but examining the effects of a reward omission on the last visit. Blue asterisks indicate significant DA decrease ($p < 0.05$, one-tailed Wilcoxon signed rank test; n = 9 rats, 1,909 sequences).

(D) Top, maze cartoon illustrating the chosen, other, and previous reward ports for an example trajectory through the maze. Bottom, multiple regression weights for the prior reward outcome at the chosen, other, and previous reward ports as effects on the DA signal (n = 9 rats, 13,448 trials; regressions performed independently for each rat; plot shows mean effect over rats). Middle, fraction of rats with significant relationships (non-zero regression coefficient, two-tailed t test) between prior reward and DA. All error bands show ±SEM.

(E) Example of one session showing trial-by-trial evolution of port (Q) values. Numbers at the top indicate nominal reward probabilities for the three ports (each in a different color to represent [top:bottom left:bottom right] ports), while tick marks indicate reward outcome on each trial (tall, rewarded; short, omission).

(F) Example value-iteration result from a single trial, spatially discounting the destination port's Q value over all hexes. Arrows point toward the destination port, and values are defined at entry into a specific hex from a specific direction.

(G) Predicted value (left) and observed DA (right) during two runs through the maze in one block from the session in (E). Top example uses the same value map as (F).

(H) Regression coefficients for hex value from a mixed-effects regression predicting hex-level DA (n = 9 rats, 77 sessions, 13,381 trials, 230,252 hex entries). Bar shows fixed effect over rats; diamonds show fixed effect for each rat over sessions.

(I) Regression model's coefficients of partial determination for value and running speed. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

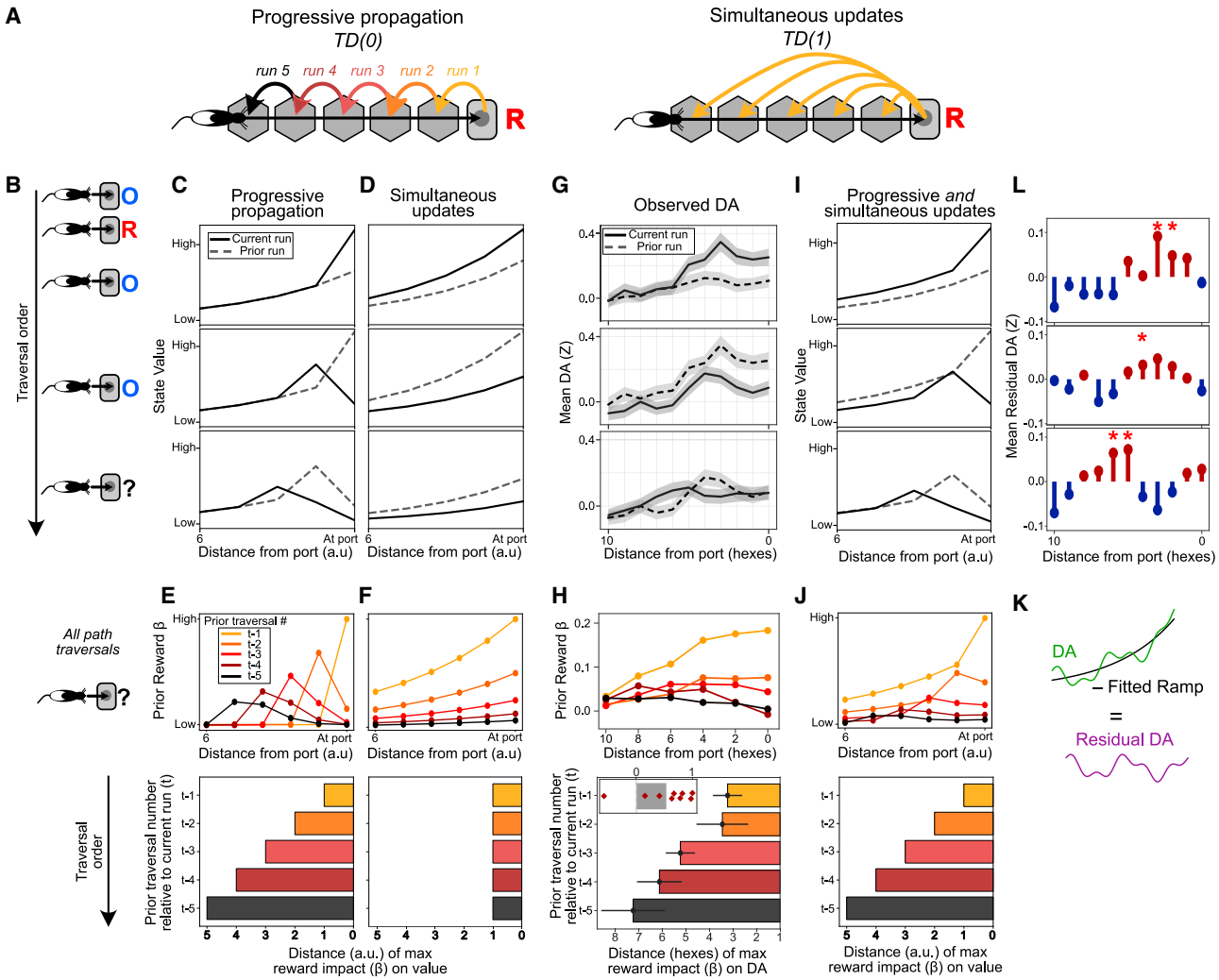


Figure 4. Progressive propagation of DA signals across space

(A) Cartoon contrasting propagating versus simultaneous value-update algorithms. Left, in TD(0), the impact on value coding of a single reward progressively moves back along the state sequence over subsequent runs along the same path. Right, in TD(1), a single reward immediately updates values for all states experienced during that trial.

(B) Illustrative outcome sequence for successive runs to the same port, with a single reward among a series of omissions.

(C) Value function from a simulated TD(0) learner over the final three traversals of the sequence in (B). Solid lines indicate the value function during the current run; dashed lines show the value function during the previous run to illustrate changes.

(D) Value function from a TD(1) learner over the same three sequential traversals.

(E and F) Analyzing the distance from the reward port at which prior rewards have their strongest impact (linear regression weight) on state value. Top, multiple regressions of state value to a path's prior five reward outcomes, at each distance. Bottom, average distance from the port where each prior reward outcome has the *strongest* effect on value. **(**E) Predictions from the TD(0) algorithm over 1,000 simulated successive traversals of the same path, with rewards delivered randomly at 50% probability. (F) Same analysis as (E), but for TD(1).

(G) Observed mean DA traces for the trial order corresponding to (C); (9 rats, n = 247 groups of trials; error bands indicate ±SEM).

(H) Results from the same analysis as in (E) and (F), but for DA over all successive traversals of the same path irrespective of reward outcome (n = 9 rats, 13,427 trials, 235,524 hexes; binned by units of two hexes for clear visualization of effect). Bar plot shows mean effect over rats ±SEM. Bottom inset, correlation between the distance of the peak reward effect on DA (in hexes) and the prior traversal number (1–5 previous traversals). Bar shows mean over rats; diamonds show individual rat coefficients (p = 0.00195, two-tailed Wilcoxon signed rank test; statistical significance is maintained over a range of hex bin sizes).

(I) Predicted value function for a linear combination of TD(0) and TD(1) value updates, over the final three traversals shown in (C).

(J) Same as (E) and (F), but for the combined TD(0) and TD(1) simulations.

(K and L) Examination of deviations from a smooth ramp. (K) Illustration of an individual-trial DA trace (green), the fitted average ramp for subtraction (black), and the remaining DA residuals for analysis. (L) Observed mean DA residuals over the same three traversals as in (C). Red asterisks indicate hexes where observed mean residual DA was higher than 95% of a shuffled null distribution at that hex (see STAR Methods).

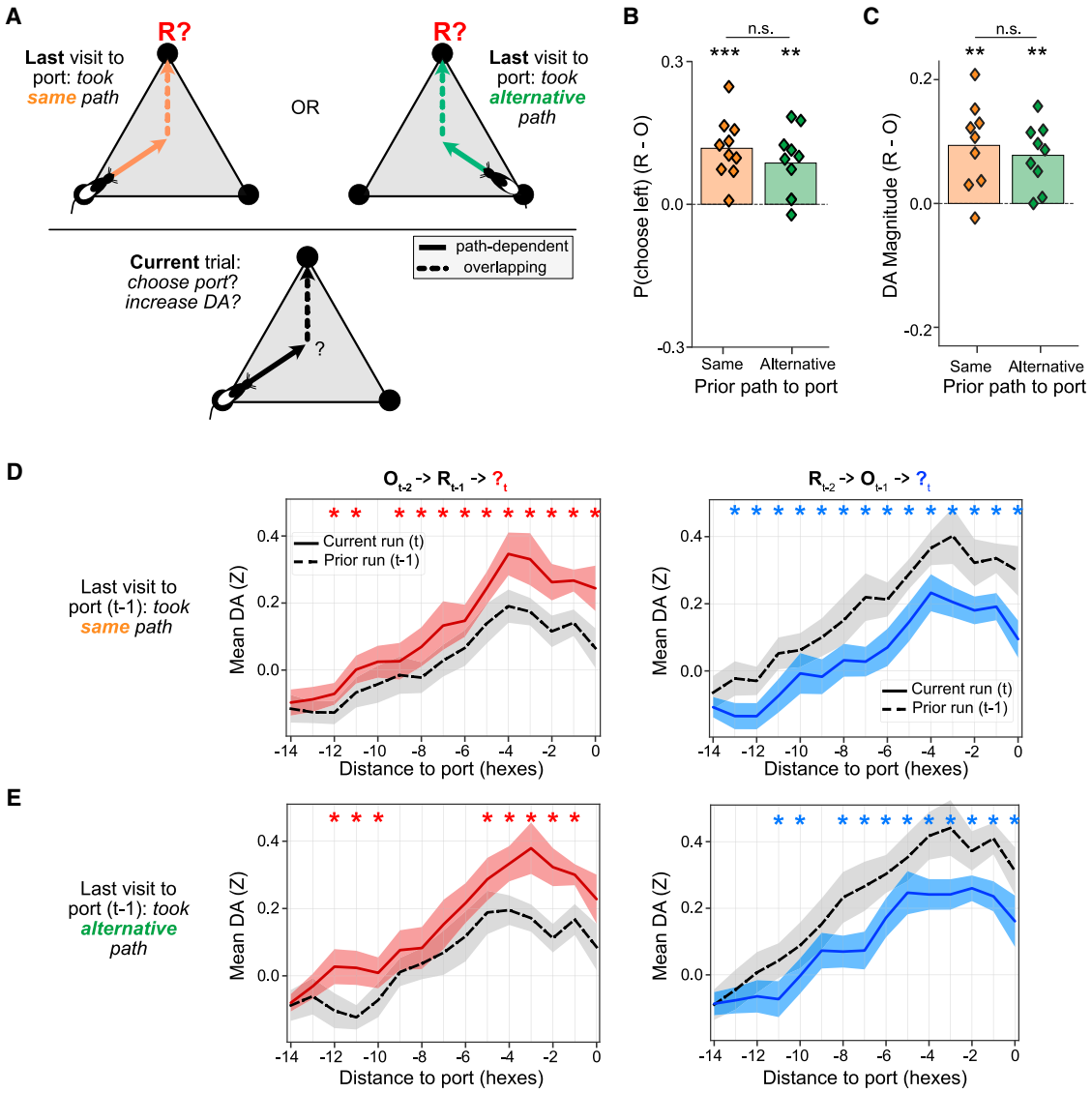


Figure 5. Model-based inference globally updates DA place values and guides choices

(A) Cartoon of two distinct routes a rat could have taken on the previous visit to a reward port (top of triangle). Portions of each route are distinct based on the starting location (path-dependent hexes; solid line), while other portions overlap (dotted line).

(B) The probability of choosing the left (counterclockwise) of the two available ports after a reward, compared with an omission. Analysis was separated by trials where, the last time that port was visited, the rat took either the same or alternative path. Bars show aggregate means, and points show individual rat values (n = 10 rats; 2,823 rewarded trials along same path, 2,439 omission trials along same path, 1,799 rewarded trials along alternative path, 1,433 omission trials along the alternative path).

(C) DA magnitude in path-dependent hexes following a reward, compared with an omission, the last time the destination port was visited from either the same (n = 9 rats, 2,500 rewarded, 1,752 omission trials) or alternative path (n = 9 rats, 1,790 rewarded, 1,337 omission trials). *p < 0.05, **p < 0.01, and ***p < 0.001. There was no difference between same and alternative conditions for either choice or DA (paired two-tailed Wilcoxon signed rank test, p = 0.275 and 0.570 respectively).

(D) DA ramps when rats previously took the same path to the destination port. Left, mean DA over successive path traversals to the same port, where reward was omitted two visits ago (t−2), but reward was delivered on the prior visit (t−1; n = 9 rats, 1,087 trials). Red asterisks indicate a significant increase in hex-level DA (p < 0.05, one-tailed Wilcoxon signed rank test). "R" and "O" denote rewards and omissions, respectively, on the t−n previous visits to the port. Right, same as left but examining the effects of a reward omission the last time the path was taken. Blue asterisks indicate a significant DA decrease (p < 0.05, one-tailed Wilcoxon signed rank test; n = 9 rats, 1,079 trials).

(E) Same as (D), but examining trials where rats previously took the alternative path to the destination port. Left, same as (D) left, but where the rat previously took the alternative path to the destination port (n = 9 rats, 592 trials). Right, same as (D) right, but where the rat previously took the alternative path to the destination port (n = 9 rats, 583 trials).

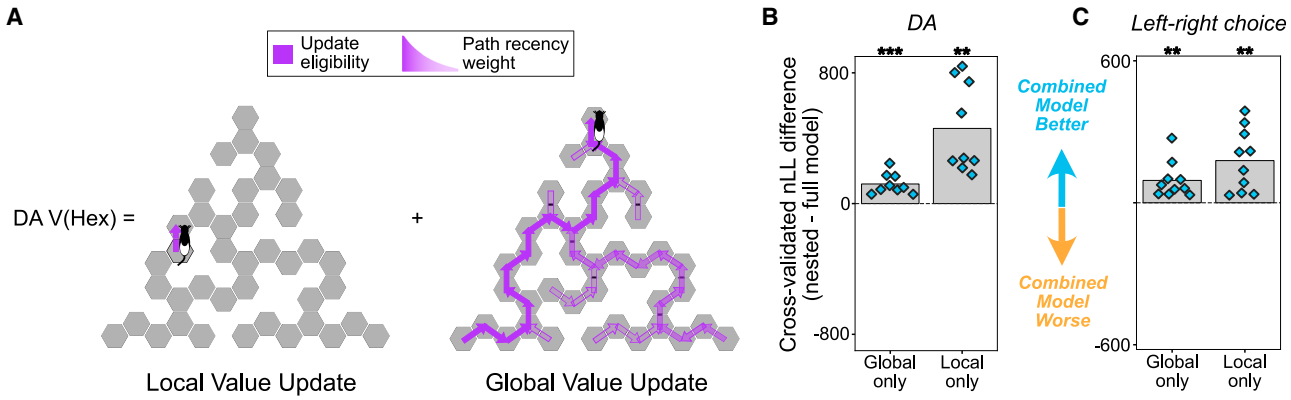


Figure 6. A combined local and global value-update model accounts for hex-level DA

(A) Illustration of the dual-component hex-value RL model's local and global learning algorithms. Left, TD(0) updates the value of the rats' experienced hex state at each hex transition. Right, upon port entry, a model of the maze's structure (the hexes that both led and could have led to the goal port) is used to globally update hex-value estimates. The model is updated each trial as a weighted sum of previously traversed paths to the chosen port; see STAR Methods for details.

(B) Gain or loss of model fit to hex-level DA when each learning component is removed from the dual-component model. Positive values indicate superior dual-component performance (negative log likelihood [nLL]), compared with single-component performance. Bars show mean fit comparison over all rats together; diamonds show model comparisons for individual rats. Removing either the TD or global components provided a significantly worse relationship to the observed DA signals (global only: $p = 0.0005$, two-tailed one-sample t test, t statistic = 5.70; local only: $p = 0.0010$, two-tailed one-sample t test, t statistic = 5.032).

(C) Same as (B) but assessing the likelihood of each left-right choice between hexes in the maze. Removing either the local or global components provided a significantly worse relationship to the observed hex-level choice behavior (global only: $p = 0.0033$, two-tailed one-sample t test, t statistic = −3.962; local only: $p = 0.0021$, two-tailed one-sample t test, t statistic = −4.28).

reward had been obtained taking the same path or the alternative (Figures 5D and 5E). Thus, NAc DA signals reflect MB calculations of inferred future reward from any location, in addition to the MF TD-like learning from direct experience.

Dual processes account for NAc DA signals during goal approach

To confirm that DA signals are best modeled as arising through the combination of MB and MF learning mechanisms, we applied a dual-process hex-level RL algorithm (Figure 6). This RL agent experienced the same sequence of hexes and rewards as each rat and generated corresponding value estimates at each moment. Upon each transition between hexes, MF TD ($\lambda = 0$) locally updated just the value of the previous hex state. The second, MB, process updated the values of *all* hexes throughout the maze each time a reward port was visited. This global update relied upon the rats' evolving knowledge of maze structure, maintained as a recency-weighted average of the tendency of each hex to be followed by a visit to each specific port (whether rewarded or not; Figure 6A; see STAR Methods). Regression analysis revealed a significant relationship between values in this dual-process model and observed DA (mean $\beta = 0.798 \pm 0.075$ SEM; $p < 0.05$ in 8/9 rats, Wald test; rat population $p = 5.408 \times 10^{-6}$, two-tailed one-sample t test, t statistic = 10.620).

We then compared the performance of the dual-process model in explaining DA signals with two nested models,[36] each with one update process removed (by setting the learning rate to zero). The combined model outperformed either process alone (Figure 6B). This result was observed consistently across individual rats and sessions (Figure S5), ruling out the possibility that individual rats use MF or MB processes idiosyncratically. Taken together, this series of analyses provides strong evidence that NAc DA reflects values that are jointly updated using two classes of learning algorithm: chained updates across sequentially traversed states and maze-wide updates through MB inference.

Finally, we assessed whether these dual credit assignment processes are used to guide rats' decisions at the level of the individual hex. Using the same dual-component hex-value RL algorithm, we tested whether a model with both learning components explained rats' hex choices at decision points better than nested models with one process removed (see STAR Methods). Parallel to the NAc DA place-value map, rat hex choices were best explained by values updated using a combination of TD(0) and MB updates (Figure 6C).

DISCUSSION

Theoretical models of RL and decision-making have very often employed multi-step navigation through simulated mazes to investigate the performance of distinct algorithms.[1] As RL models form the standard framework for interpreting DA signals, it is perhaps surprising that the present study is the first—to our knowledge—to examine real-time DA dynamics in a rich and dynamic spatial environment.

Our observation of a DA pulse at reward receipt, scaling with positive RPE, is consistent with standard DA ideas (although it is noteworthy that such pulses were not observed in a prior study using a simpler T-maze[19]). By contrast, we did not expect to see a similarly sized DA pulse when rats detected a newly opened path. It is natural to interpret this as some form of error signal, but it is not yet clear what type. DA signals have long been associated with novelty and salient events,[37,38] and some theories

have argued that DA can signal a range of prediction errors beyond simply RPE.[39,40] However, a newly blocked path did not evoke a comparably sized DA pulse, suggesting that the relevant feature is not simply an unexpected stimulus or the need to update models of the environment (as in successor representations[40]). It appears that the DA pulse is related to the newly discovered opportunity for action,[41] perhaps reflecting the value of discovering possible new paths to reward through exploration.[42,43] Specifying the underlying information processing in greater detail will require further experiments with a larger dataset of barrier movements.

A major objective of this study was to investigate the ramps in NAc DA release that occur as unrestrained animals approach rewards.[5,16–18] We and others have previously interpreted this ramping DA as reflecting increasing reward expectation (a.k.a. value). Consistent with this, we found here that ramps track the animals' recent reward history, for example, ramping more strongly when the destination port was rewarded at the last visit. As rats ran through the maze, the moment-by-moment DA levels formed a dynamic map of values: expected rewards discounted by distance from the reward port. This result contributes to an ongoing discussion about whether/how DA signals reflect the costs, as well as the benefits, of potential decisions.[44] In the maze, rats clearly treated distance as a cost, as shown by their reluctance to choose paths leading to more distant reward ports. This cost was incorporated into the DA signal through spatial discounting, producing a net value signal potentially useful for governing motivation from each point. This interpretation fits well with observations that lesions of NAc DA shift motivation in cost/benefit decision-making in maze tasks[45,46] and that boosting DA can immediately enhance motivation to work.[17] Consistent with this interpretation, instantaneous NAc DA was predictive of speed shortly afterward, as if energizing effort in the pursuit of reward. An interesting area for future studies is how discounting future rewards over space relates to discounting over time, which has been previously reported for DA signals[47] and may involve distinct timescales in different striatal subregions.[48]

Alternative accounts have emerged arguing that DA ramps reflect RPE. This is possible under various assumptions: e.g., that values are rapidly forgotten,[20] that there are constraints on the functional form by which value decays in space or time,[49] or that animals are uncertain about their current state.[50] The present study was not specifically designed to test those ideas. Nonetheless, the strong correspondence between DA dynamics and upcoming reward estimates observed here makes a value interpretation of ramps the most parsimonious. This value coding would be separate to the RPE coding noted above (although a recent study has argued that ramps themselves may contribute to credit assignment[51]). Evidence that DA ramping can be mechanistically distinct from RPE coding comes from a prior study in which we compared DA cell firing with release.[5] Discrete reward cues evoked RPE-encoding burst firing of identified ventral tegmental area (VTA) DA cells and a parallel increase in NAc DA release. By contrast, NAc DA ramps appeared to occur even without any increase in DA cell firing, suggesting a separate process. A similar comparison in the context of our maze task could be highly illuminating.

Regardless of whether ramping DA signals value in addition to or as a side effect of error signaling, our results provide new insights into the specific algorithms by which these signals are updated. TD learning has been central to models of DA signals for decades,[2] but evidence for the signature progressive backward propagation of value over trials has been limited.[7–9] Using DA as a readout of values, we clearly observed just such propagation of value across space. This observation may have been aided by our maze design, in which each hex can correspond to a discrete left/right decision point, and may thus be more likely to be treated by the brain as a distinct "state." Nonetheless, we could not force the rats to treat hexes as states, and indeed inspection of the propagating DA "bump" suggests that the actual spatial resolution employed by the internal TD algorithm may be in the range of ~2–3 hexes (Figure 4).

Although TD learning is visibly present, we also demonstrated that rats additionally assign credit over long distances in a single step and over paths not directly experienced, suggesting they employ internal models of their environment to guide their DA signals and foraging decisions. We cannot currently say, however, exactly *when* they are using such models. For simplicity, we simulated MB value updates as occurring when outcomes are revealed at reward ports. This may be the right time: after running along trajectories through mazes, rats often show sharp-wave ripple (SWR) events in which dorsal hippocampal place cells can replay recently taken trajectories.[52] This replay is especially common after reward receipt[53,54] and has been proposed to update values along the encoded trajectories.[52] Replay can also encode alternative potential paths to reward,[11,55,56] providing a potential mechanism underlying MB inference of updated values.[57] Echoing this perspective, recent research in AI has increasingly emphasized the use of models retrospectively for credit assignment.[58,59] Credit assignment might involve synaptic plasticity downstream of reactivated place representations[60] and/or reconfiguration of network dynamics, which can support fast adjustments in value-guided decision-making.[61]

However, MB value updates might instead, or additionally, be occurring in a prospective manner as rats run toward goals (similar to another family of AI algorithms that that use models for planning[62]). First, SWRs occur not only following reward receipt but also during pauses in behavior, when place cells can encode locations predictive of the animal's future path.[63] Such "forward" replay toward goal locations could, in principle, accomplish MB value updates similar to "backward" replay from them.[57] Additionally, actively running rats show "theta sequences" in which the maze location encoded by hippocampal place cells sweeps forward ahead of the animal within each theta cycle[64] and can even rapidly switch between representations of distinct possible future paths.[65,66] It may be that theta sequences help retrieve current values of potential goal locations to help guide decision-making, although this is not yet known. We note that such cognitively demanding planning processes may be only activated when the brain perceives a need to do so. If DA ramping is linked to prospective planning, this could explain why DA ramps peak just ahead of actually reaching the goal (Figure 3). Within the last three hexes of the path the reward port is directly visible, so no internal calculations are required for navigation. This account is also consistent with prior reports

that ramps disappear entirely as rat behavior becomes increasingly routine,[18,22] and they can reappear immediately if task contingencies change.[18] There is analogous evidence that non-local activity in hippocampus declines over repeated experience, including both SWRs[54] and cycling between multiple paths during theta sequences.[67] Furthermore, there is substantial behavioral and pharmacological evidence that NAc DA is specifically required when animals need to flexibly calculate trajectories to reward (e.g., from a variable start location) rather than performing a stereotyped sequence of actions.[68,69] For future reports, we aim to combine measurements of DA ramps with high-density hippocampal recordings to gain greater access to the internal calculations driving DA dynamics during active foraging.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Animals
- METHODS DETAILS
  - Behavioral task
  - Fiber photometry
  - Reinforcement learning models
  - Data Analyses
- QUANTIFICATION AND STATISTICAL ANALYSES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2023.07.017.

ACKNOWLEDGMENTS

We thank Simon Little, Colin Hoy, Vijay Namboodiri, Anna Grzymala-Busse, and members of the Berke Lab for providing valuable feedback on manuscript drafts; Ali Mohebi for initial assistance with fiber photometry procedures, hardware configuration, and surgical procedures; Yang-Sun Hwang for assistance with rat training and fiber photometry; Lily Pelattini for assistance with histology; and other members of the Berke Lab for assistance and advice. This work was supported by the National Institute on Drug Abuse, the National Institute of Neurological Disorders and Stroke, the National Institute of Mental Health, the State of California, and the University of California, San Francisco.

AUTHOR CONTRIBUTIONS

T.A.K. developed, built, and optimized the maze task; performed the photometry experiments; analyzed the photometry data; and performed the computational modeling in partnership with N.D.D. A.E.C. and L.M.F. provided regular scientific input and feedback. A.E.K. performed RL modeling of hex-level choice behavior. N.D.D. provided guidance and code for computational modeling and developed the computational models with T.A.K. J.D.B. designed and supervised the study, developed the maze task, and wrote the manuscript together with T.A.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 8, 2023
Revised: June 23, 2023
Accepted: July 25, 2023
Published: August 22, 2023

REFERENCES

1. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction, Second Edition (MIT Press).
2. Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.
3. Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141. https://doi.org/10.1016/j.neuron.2005.05.020.
4. 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. https://doi.org/10.1038/nature10754.
5. Mohebi, A., Pettibone, J.R., Hamid, A.A., Wong, J.-M.T., Vinson, L.T., Patriarchi, T., Tian, L., Kennedy, R.T., and Berke, J.D. (2019). Dissociable dopamine dynamics for learning and motivation. Nature *570*, 65–70. https://doi.org/10.1038/s41586-019-1235-y.
6. Hart, A.S., Rutledge, R.B., Glimcher, P.W., and Phillips, P.E.M. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J. Neurosci. *34*, 698–704. https://doi.org/10.1523/JNEUROSCI.2489-13.2014.
7. Pan, W.X., Schmidt, R., Wickens, J.R., and Hyland, B.I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci. *25*, 6235–6242. https://doi.org/10.1523/JNEUROSCI.1478-05.2005.
8. Amo, R., Matias, S., Yamanaka, A., Tanaka, K.F., Uchida, N., and Watabe-Uchida, M. (2022). A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. Nat. Neurosci. *25*, 1082–1092. https://doi.org/10.1038/s41593-022-01109-2.
9. Jeong, H., Taylor, A., Floeder, J.R., Lohmann, M., Mihalas, S., Wu, B., Zhou, M., Burke, D.A., and Namboodiri, V.M.K. (2022). Mesolimbic dopamine release conveys causal associations. Science *378*, eabq6740. https://doi.org/10.1126/science.abq6740.
10. Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711. https://doi.org/10.1038/nn1560.
11. Liu, Y., Mattar, M.G., Behrens, T.E.J., Daw, N.D., and Dolan, R.J. (2021). Experience replay is associated with efficient nonlocal learning. Science *372*. https://doi.org/10.1126/science.abf1357.
12. Sharpe, M.J., Batchelor, H.M., Mueller, L.E., Yun Chang, C., Maes, E.J.P., Niv, Y., and Schoenbaum, G. (2020). Dopamine transients do not act as model-free prediction errors during associative learning. Nat. Commun. *11*, 106. https://doi.org/10.1038/s41467-019-13953-1.
13. Sadacca, B.F., Jones, J.L., and Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. eLife *5*, 1–13. https://doi.org/10.7554/eLife.13665.
14. Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. Neuron *41*, 269–280. https://doi.org/10.1016/s0896-6273(03)00869-9.
15. Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027.
16. Roitman, M.F., Stuber, G.D., Phillips, P.E.M., Wightman, R.M., and Carelli, R.M. (2004). Dopamine operates as a subsecond modulator of

food seeking. J. Neurosci. *24*, 1265–1271. https://doi.org/10.1523/JNEUROSCI.3823-03.2004.

17. Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. Nat. Neurosci. *19*, 117–126. https://doi.org/10.1038/nn.4173.
18. Collins, A.L., Greenfield, V.Y., Bye, J.K., Linker, K.E., Wang, A.S., and Wassum, K.M. (2016). Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. Sci. Rep. *6*, 20231. https://doi.org/10.1038/srep20231.
19. Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E.M., and Graybiel, A.M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature *500*, 575–579. https://doi.org/10.1038/nature12475.
20. Morita, K., and Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front. Neural Circuits *8*, 36. https://doi.org/10.3389/fncir.2014.00036.
21. Kim, H.R., Malik, A.N., Mikhael, J.G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S.J., et al. (2020). A unified framework for dopamine signals across timescales. Cell *183*, 1600–1616.e25. https://doi.org/10.1016/j.cell.2020.11.013.
22. Guru, A., Seo, C., Post, R.J., Kullakanda, D.S., Schaffer, J.A., and Warden, M.R. (2020). Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. Preprint at bioRxiv. https://doi.org/10.1101/2020.05.21.108886.
23. Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. *9*, 1057–1063. https://doi.org/10.1038/nn1743.
24. Parker, N.F., Cameron, C.M., Taliaferro, J.P., Lee, J., Choi, J.Y., Davidson, T.J., Daw, N.D., and Witten, I.B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat. Neurosci. *19*, 845–854. https://doi.org/10.1038/nn.4287.
25. Namboodiri, V.M.K. (2022). How do real animals account for the passage of time during associative learning? Behav. Neurosci. *136*, 383–391. https://doi.org/10.1037/bne0000516.
26. Foster, D.J., Morris, R.G.M., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus *10*, 1–16. https://doi.org/10.1002/(SICI)1098-1063(2000)10:1<1::AID-HIPO1>3.0.CO;2-1.
27. Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H.J., Ornelas, S., Koay, S.A., Thiberge, S.Y., Daw, N.D., Tank, D.W., et al. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. Nature *570*, 509–513. https://doi.org/10.1038/s41586-019-1261-9.
28. Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. Science *310*, 1338–1340. https://doi.org/10.1126/science.1115233.
29. Lau, B., and Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. J. Exp. Anal. Behav. *84*, 555–579. https://doi.org/10.1901/jeab.2005.110-04.
30. Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879. https://doi.org/10.1038/nature04766.
31. Huh, N., Jo, S., Kim, H., Sul, J.H., and Jung, M.W. (2009). Model-based reinforcement learning under concurrent schedules of reinforcement in rodents. Learn. Mem. *16*, 315–323. https://doi.org/10.1101/lm.1295509.
32. Patriarchi, T., Cho, J.R., Merten, K., Howe, M.W., Marley, A., Xiong, W.H., Folk, R.W., Broussard, G.J., Liang, R., Jang, M.J., et al. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. Science *360*, eaat4422.
33. Gadagkar, V., Puzerey, P.A., Chen, R., Baird-Daniel, E., Farhang, A.R., and Goldberg, J.H. (2016). Dopamine neurons encode performance error in singing birds. Science *354*, 1278–1282. https://doi.org/10.1126/science.aah6837.
34. Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology *191*, 507–520. https://doi.org/10.1007/s00213-006-0502-4.
35. Simon, D.A., and Daw, N.D. (2011). Neural correlates of forward planning in a spatial decision task in humans. J. Neurosci. *31*, 5526–5539. https://doi.org/10.1523/JNEUROSCI.4647-10.2011.
36. Daw, N.D. (2011). Trial-by-trial data analysis using computational models. In Decision Making, Affect, and Learning: Attention and Performance XXIII (Oxford University Press).
37. Horvitz, J.C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience *96*, 651–656. https://doi.org/10.1016/s0306-4522(00)00019-1.
38. Bromberg-Martin, E.S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. Neuron *68*, 815–834. https://doi.org/10.1016/j.neuron.2010.11.022.
39. Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? Nat. Rev. Neurosci. *7*, 967–975. https://doi.org/10.1038/nrn2022.
40. Gardner, M.P.H., Schoenbaum, G., and Gershman, S.J. (2018). Rethinking dopamine as generalized prediction error. Proc. Biol. Sci. *285*, 20181645. https://doi.org/10.1098/rspb.2018.1645.
41. Syed, E.C.J., Grima, L.L., Magill, P.J., Bogacz, R., Brown, P., and Walton, M.E. (2016). Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat. Neurosci. *19*, 34–36. https://doi.org/10.1038/nn.4187.
42. Agrawal, M., Mattar, M.G., Cohen, J.D., and Daw, N.D. (2022). The temporal dynamics of opportunity costs: a normative account of cognitive fatigue and boredom. Psychol. Rev. *129*, 564–585. https://doi.org/10.1037/rev0000309.
43. Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped DQN. Preprint at arXiv. https://doi.org/10.48550/arXiv.1602.04621.
44. Walton, M.E., and Bouret, S. (2018). What is the relationship between dopamine and effort? Trends Neurosci. *42*, 79–91. https://doi.org/10.1016/j.tins.2018.10.001.
45. Salamone, J.D., Cousins, M.S., and Bucher, S. (1994). Anhedonia or anergia? Effects of haloperidol and nucleus accumbens dopamine depletion on instrumental response selection in a T-maze cost/benefit procedure. Behav. Brain Res. *65*, 221–229.
46. Cousins, M.S., Atherton, A., Turner, L., and Salamone, J.D. (1996). Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. Behav. Brain Res. *74*, 189–197. https://doi.org/10.1016/0166-4328(95)00151-4.
47. Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. J. Neurosci. *28*, 7837–7846. https://doi.org/10.1523/JNEUROSCI.1600-08.2008.
48. Wei, W., Mohebi, A., and Berke, J.D. (2022). A spectrum of time horizons for dopamine signals. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.31.466705.
49. Gershman, S.J., Moustafa, A.A., and Ludvig, E.A. (2014). Time representation in reinforcement learning models of the basal ganglia. Front. Comput. Neurosci. *7*, 194. https://doi.org/10.3389/fncom.2013.00194.
50. Mikhael, J.G., Kim, H.R., Uchida, N., and Gershman, S.J. (2022). The role of state uncertainty in the dynamics of dopamine. Curr. Biol. *32*, 1077–1087.e9. https://doi.org/10.1016/j.cub.2022.01.025.
51. Hamid, A.A., Frank, M.J., Moore, C.I., Hamid, A.A., Frank, M.J., and Moore, C.I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. Cell, 2733–2749.e16. https://doi.org/10.1016/j.cell.2021.03.046.

52. Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature *440*, 680–683. https://doi.org/10.1038/nature04587.

53. Singer, A.C., and Frank, L.M. (2009). Rewarded outcomes enhance reactivation of experience in the hippocampus. Neuron *64*, 910–921. https://doi.org/10.1016/j.neuron.2009.11.016.

54. Ambrose, R.E., Pfeiffer, B.E., and Foster, D.J. (2016). Reverse replay of hippocampal place cells is uniquely modulated by changing reward. Neuron *91*, 1124–1136. https://doi.org/10.1016/j.neuron.2016.07.047.

55. Barron, H.C., Reeve, H.M., Koolschijn, R.S., Perestenko, P.V., Shpektor, A., Nili, H., Rothaermel, R., Campo-Urriza, N., O'Reilly, J.X., Bannerman, D.M., et al. (2020). Neuronal computation underlying inferential reasoning in humans and mice. Cell *183*, 228–243.e21. https://doi.org/10.1016/j.cell.2020.08.035.

56. Bhattarai, B., Lee, J.W., and Jung, M.W. (2020). Distinct effects of reward and navigation history on hippocampal forward and reverse replays. Proc. Natl. Acad. Sci. USA *117*, 689–697. https://doi.org/10.1073/pnas.1912533117.

57. Mattar, M.G., and Daw, N.D. (2018). Prioritized memory access explains planning and hippocampal replay. Nat. Neurosci. *21*, 1609–1617. https://doi.org/10.1038/s41593-018-0232-z.

58. van Hasselt, H., Madjiheurem, S., Hessel, M., Silver, D., Barreto, A., and Borsa, D. (2021). Expected eligibility traces. Proceedings of the AAAI Conference on Artificial Intelligence*35*, pp. 9997–10005. https://doi.org/10.1609/aaai.v35i11.17200.

59. Harutyunyan, A., Dabney, W., Mesnard, T., Azar, M., Piot, B., Heess, N., van Hasselt, H., Wayne, G., Singh, S., Precup, D., et al. (2019). Hindsight credit assignment. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.02503.

60. McNamara, C.G., Tejero-Cantero, Á., Trouche, S., Campo-Urriza, N., and Dupret, D. (2014). Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. Nat. Neurosci. *17*, 1658–1660. https://doi.org/10.1038/nn.3843.

61. Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. *21*, 860–868. https://doi.org/10.1038/s41593-018-0147-8.

62. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489. https://doi.org/10.1038/nature16961.

63. Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. Nature *497*, 74–79. https://doi.org/10.1038/nature12112.

64. Wikenheiser, A.M., and Redish, A.D. (2015). Hippocampal theta sequences reflect current goals. Nat. Neurosci. *18*, 289–294. https://doi.org/10.1038/nn.3909.

65. Kay, K., Chung, J.E., Sosa, M., Schor, J.S., Karlsson, M.P., Larkin, M.C., Liu, D.F., and Frank, L.M. (2020). Constant sub-second cycling between representations of possible futures in the hippocampus. Cell *180*, 552–567.e25. https://doi.org/10.1016/j.cell.2020.01.014.

66. Comrie, A.E., Frank, L.M., and Kay, K. (2022). Imagination as a fundamental function of the hippocampus. Philos. Trans. R. Soc. Lond. B Biol. Sci. *377*, 20210336. https://doi.org/10.1098/rstb.2021.0336.

67. Johnson, A., and Redish, A.D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. *27*, 12176–12189. https://doi.org/10.1523/JNEUROSCI.3761-07.2007.

68. Nicola, S.M. (2010). The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. J. Neurosci. *30*, 16585–16600. https://doi.org/10.1523/JNEUROSCI.3958-10.2010.

69. Ikemoto, S., and Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Res. Brain Res. Rev. *31*, 6–41. https://doi.org/10.1016/s0165-0173(99)00023-5.

70. Martianova, E., Aronson, S., and Proulx, C.D. (2019). Multi-fiber photometry to record neural activity in freely-moving animals. J. Vis. Exp. 1–9. https://doi.org/10.3791/60278.

71. Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., and Mathis, M.W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat. Protoc. *14*, 2152–2176. https://doi.org/10.1038/s41596-019-0176-0.

72. Pitis, S. (2018). Source traces for temporal difference learning. Proceedings of the AAAI Conference on Artificial Intelligence*32*. https://doi.org/10.1609/aaai.v32i1.11813.

73. Huys, Q.J.M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R.J., and Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. PLoS Comput. Biol. *7*, e1002028. https://doi.org/10.1371/journal.pcbi.1002028.

74. Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. J. R. Stat. Soc. B *61*, 479–482. https://doi.org/10.1111/1467-9868.00188.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-Calbindin	Swant	Cat# CB38a
anti-GFP	Abcam	Cat# ab13970
Bacterial and virus strains		
AAVDJ-CAG-dLight1.3b	Vigene	N/A
Software and algorithms		
Custom code	Berke lab	https://doi.org/10.5281/zenodo.8172780
Expectation Maximization code	Daw lab	https://github.com/ndawlab/em
AirPLS photometry processing algorithm	Martianova et al.[70]	https://github.com/katemartian/Photometry_data_processing
LabView	National Instruments	https://www.ni.com/en-us.html
Bonsai	Bonsai Foundation	https://bonsai-rx.org/
Matlab	Mathworks	https://www.mathworks.com/products/matlab.html
Deposited data		
Processed (isosbestic-controlled) photometry and behavioral data	Mendeley Data	https://doi.org/10.17632/m59zdjpm9h.1

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data and code should be directed to and will be fulfilled by the lead contact, Josh Berke (joshua.berke@ucsf.edu).

Materials availability

The study did not generate new unique reagents.

Data and code availability

- Photometry and behavioral data are deposited at Mendeley Data: https://doi.org/10.17632/m59zdjpm9h.1
- Original code is deposited to a public lab-maintained GitHub repository: https://github.com/Berke-lab/DA_maze, and registered on Zenodo: https://doi.org/10.5281/zenodo.8172780
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Animals

All animal procedures were approved by University of California San Francisco Institutional Committees on Use and Care of Animals. Male (300–650g) and female (250-400g) wild-type Long-Evans rats (4-10 months old, bred in house) were maintained on a reverse 12:12 light:dark cycle and tested during the dark phase. Rats were mildly water deprived, receiving 30 minutes of free water access daily in addition to fluid rewards earned during task performance. During water deprivation, rat weights were maintained above 85% their baseline weight.

METHODS DETAILS

Behavioral task

The maze consists of a 1.30m-per-side equilateral triangular platform with liquid reward ports at each vertex. Solenoid valves control delivery of sucrose solution (10% sucrose, 0.1% NaCl) in 15μL droplets. Infrared photobeam sensors detect entry into the reward

ports. To prevent uncertainty over reward delivery, a brief (70ms) 3.0 kHz tone was played through a speaker below the center of the maze immediately before solenoid valve opening. Equally spaced columnar barriers divide the maze into 49 hexagonal units (“hexes”). Additional barriers can be placed in any combination of the 49 hexes to create unique maze configurations. The apparatus was controlled by an Arduino Mega, while the Open Ephys software, Bonsai, was used for behavioral and video data acquisition.

Prior to implantation, rats were mildly water deprived and trained in the maze for approximately three weeks. Pre-training consisted of learning to poke into reward ports to receive reward, at 100% delivery probability with no additional barriers. Rats were pretrained until they completed an average of at least one trial per minute in a 60-minute session (1-2 sessions to reach criterion on average). To discourage a sit-and-wait strategy, after each visit to a port that port was not rewarded again until another port was visited (this rule is present throughout training and testing). Rats were then trained on the task until reaching criterion (>= one trial per minute in a 90 to 120-minute session).

Before each session, barriers (8 or 9) are added to the maze to create a configuration that is novel to the rat. To prevent clearly visible paths between ports, we ensured that at least one barrier obstructed each direct path. We also configured at least one path to be longer or shorter than another path, to create distinct distance costs associated with different paths. In the probability-change variant, the maze configuration stays consistent throughout a session, but reward probabilities are changed following each block (50-70 trials). Probabilities are reassigned pseudo-randomly, according to the rule that the most rewarding port and the least rewarding port are not the same for two consecutive blocks. In the barrier-change variant of the task, reward probabilities remain fixed throughout the session, while one barrier is moved at each transition between blocks. Upon a block transition, barriers are moved strategically to simultaneously alter the lengths of multiple paths: at least one path will increase in length, and at least one will decrease in length. Critically, a short path prior to the block change does not necessarily become long afterwards, making it impossible for the rat to make inferences about which paths have become longer and shorter. Barriers were physically moved by the experimenter, who entered the task area after the rat poked into a reward port on the last trial of a block. To prevent the development of associations between experimenter entry and configuration changes, the experimenter randomly entered the task area to briefly raise and lower a barrier – without changing the maze configuration – at least once during each training session. Each daily test session used either the probability-change or barrier-change variant, and we only included behavioral sessions with 100 or more trials for further analysis. Individual rats were also excluded from analysis altogether if logistic multiple regression revealed a non-significant effect of either reward probability or distance cost on their port choices (n=1 rat without significant reward effect, and n=1 rat without significant distance effect from a total initial dataset of n = 12 rats). All rats experienced sessions with port-reward probabilities drawn from a set of [0.9, 0.5, 0.1], but four rats also had probabilities drawn from [0,8, 0.5, 0.2] on a subset of sessions.

Rats’ implant caps were labeled and tracked using Deeplabcut.[71] Custom code was used to segment the maze into hexes and classify hex occupancy. For time points with missing position information (i.e., when rat’s heads were momentarily obstructed by barriers), we used the maze’s hex adjacency matrices to interpolate between hexes.

Fiber photometry

The nucleus accumbens core was bilaterally targeted using the following coordinates in relation to bregma: +/-1.7mm medial, 1.7mm anterior, and 6.2mm below brain surface. Virus – 1μL of AAVDJ-CAG-dLight1.3b (Vigene) at a titer of $2x10^{12}$ – was delivered using a stereotaxic injection pump (Nanoject III). Virus was injected 200μm ventral to the target coordinates, as described in Mohebi et al.[5] During the same surgery, 200μm optical cannulae were subsequently implanted and cemented in place. A subset of rats (n = 4; IM-1322, IM-1398, IM-1434, IM-1478), were also implanted with a custom electrophysiology probe in the dorsal hippocampus.

Rats were removed from water deprivation at least 24 hours prior to surgery. One week after surgery, rats began mild water deprivation and were retrained on the task, while waiting for expression of dLight. Rats began photometry recordings in the maze at least two full weeks following surgery. Only one implanted fiber was recorded in a given photometry session.

Photometry data acquisition methods have been described previously.[5] Baseline correction was performed using the adaptive iteratively reweighed Penalized Least Squares (airPLS) algorithm.[70] Baseline-subtracted 470nm and 405nm (isosbestic control) signals were then each standardized (z-scored) using a session-wide median and standard deviation. The standardized reference signal was fitted to the 470nm using non-negative robust linear regression, and the normalized fluorescence signal was computed by subtracting the fitted reference signal from the standardized dLight signal. To reduce the frequency and severity of optical artifacts, we used a pigtailed optical commutator (Doric Lenses), oriented horizontally, and manually controlled its movement using a custom stepper-motor interface. Recording locations were histologically verified using immunohistochemistry.[5] Recording sessions were excluded if a recording failure occurred at any point during the session, such as an optical fiber becoming broken or unplugged.

For all time-based analyses, the dLight signal was downsampled to 250 Hz and smoothed with a rectangular 100 ms rolling mean. For hex-level photometry analyses, we calculated the mean dopamine within each traversed hex on a given run. For comparison with RL model variables, we computed mean dopamine within each traversed hex from each possible direction of entry. This included repeat entries into hexes traversed multiple times within a trial (e.g., after leaving a hex, entering a dead end, and running back to through that same hex). To avoid analyzing subsets of data where rats mistakenly returned to the previous port (where reward is unavailable), only data between the final poke at one port and the first poke at a different port were included. Distance from the destination port was computed as the shortest possible distance, in hexes, to the destination port from the current hex, according to the current maze map. For event-aligned plots, traces were first averaged over sessions within each rat before taking the average over

each rat, unless otherwise specified. Unless otherwise specified, we treated individual rats as the unit of analysis, rather than e.g. fiber recording locations.

Reinforcement learning models

Q(port) learning

To estimate the rats' expected value at each port on each trial, we used a simple, trial-based Q learning algorithm. The model learns values associated with each port using the following update rule:

$$Q(port_t) \leftarrow (1 - \alpha)Q(port_t) + \alpha R_t,$$

where α is the learning rate, t denotes the current trial, and R denotes reward received at the end of the trial. Choice was modeled as a probabilistic decision between the two available destination ports, left ("L") and right ("R"), denoted by their position clockwise or counterclockwise from the animal, on each trial using a softmax distribution:

$$P(c_t = c \in L,R) \propto \exp(\beta Q(c) + \beta_{ccw} IsLeft(c) + \beta_{dist}(dist(c)))$$

The inverse temperature parameter, β, controlled the degree to which the value of the destination port, $Q(port)$ influenced choice. The ("β_{ccw}") term was added to control for leftward (counterclockwise) turn biases, and a distance-sensitivity ("β_{dist}") term was added to control for effort cost scaling with the distance $dist(c)$ to the port. "IsLeft" encodes whether the choice, "c", was leftward from the current port. Parameters were optimized to maximize fit to rats' observed port choices.

Value iteration

We sought to generate spatially discounted chosen value estimates for each hex at the individual-trial level, in a manner faithful to the maze configuration on each trial. We first specified ground truth hex-state transition matrices for each unique maze configuration. We then used a value-iteration[1,35] algorithm to dynamically estimate state value over each hex-state. Here, hex-states were defined by hex ID (1-49) paired with the direction of hex entry, which resulted in a 126-hex-state state space (each hex has between one and three possible directions of entry). For each trial, the reward function was set to zero at all states other than the chosen port, which was set to the goal port's Q value on that trial. Hex values were initialized at zero, and value was iteratively learned by taking the maximum of the available discounted next-state values, over all hexes, until convergence. The update rule took the following form:

$$V(state) \leftarrow \max_{a \in (L,R)} (\gamma V(nextstate(state, a))) \text{ for all hex-states state}$$

where "a" is a left or right exit from the current hex-state, and $nextstate(state, a)$ is the state obtained (through the transition matrix) by exiting *state* with action *a*. The discount factor, γ, was optimized for each behavioral session to maximize the fit to DA (minimizing negative log likelihood of the observed DA, given the estimated value function[36]).

TD(λ) toy-path value learner

To test distinct predictions about reward propagation over space, we created a simple TD model with an adjustable eligibility-trace parameter (TD(λ) with replacing traces[1]). Each traversed state was associated with an update eligibility that decayed exponentially – by a factor of λ – with each timestep (state transition). To model locally chained value propagation, we implemented a one-step TD model by setting λ equal to zero (TD(0)). To model updating over the entire traversed path, we set λ equal to one. Due to the absence of RPE during successive traversals of the same path under TD(1), value updates only occur at the terminal state, and for the entirety of the traversed path. Under these conditions, TD(1) is equivalent to a Monte-Carlo learning process.[1] Eligibility traces e were initialized at zero, and the update rules were as follows, at each step t:

$$e(state) \leftarrow \lambda\gamma e(state) \text{ for all states}$$

$$At\ non\text{-}port\ states : e(state_t) \leftarrow 1$$

$$\delta_t = R_t + \gamma V(state_{t+1}) - V(state_t)$$

$$V(state) \leftarrow V(state) + \alpha e(state)\delta_t \text{ for all states}$$

$$Upon\ reaching\ port : e(state) \leftarrow 0 \text{ for all states}$$

where V is the value function, γ is the discount factor, and α is the learning rate. For clear visualization of model predictions, TD(0) α was set to 0.85 and γ was set to 0.8; TD(1) α was set to 0.5 and γ was set to 0.8 . To recreate a ramp similar to the DA signal, each learner started with a baseline value function peaking at 0.4 and discounted by a factor of 0.8. The toy environment was implemented as a six-state sequential path to a reward port, and the reward function equaled zero at all states except the terminal port. Port reward

sequences could be set by the experimenter in order to visualize the resulting value functions. Alternatively, rewards could be drawn from a random distribution. For the regression analysis in Figure 4, assessing the relationship between prior reward outcomes and model value estimates at each state, we simulated 1000 trials with random rewards delivered at 50% probability. To illustrate one possible combination of TD(0) and TD(1) learners (Figures 4I and 4J), we took a weighted sum of the outputs of each (choosing weights of 0.3 and 0.7 respectively, without any fitting).

Dual-component hex-value learner

To compare contributions of spatially local TD and maze-wide inference-based learning processes, we developed a value learning algorithm over hex-states (location and direction, defined as before), with two separate value-update components: local TD(0) value learning, and a maze-wide model-based update.

A one-step TD(0) update occurred at every hex entry according to the following update rule:

$$\textit{Learning rule at each hex transition}: V(state_t) \leftarrow V(state_t) + \alpha_{TD}(\gamma_{MF} V(state_{t+1}) - V(state_t))$$

$$\textit{Learning rule upon port entry}: V(state_t) \leftarrow V(state_t) + \alpha_{TD}(R_t - V(state_t))$$

where α_{TD} is the TD learning rate, and γ_{MF} is the spatial discount factor. The reward function, R, was zero for all non-port hexes. Hex-state values were initialized at 0.2, to convey a small uniform expectation of future reward from all locations. Upon reaching a reward port, model-based updates were also performed over the entire map according to the following rule:

$$V(state) \leftarrow V(state) + \alpha_{MB} T(port_t, state)(R_t - V(state)) \text{ for all states,}$$

where α_{MB} is the model-based update learning rate, and $T(port, state)$ weights the update by the discounted on-policy distance from each state to the current port. This map is learned online by recency-weighted averaging over states encountered on paths into the port. In particular upon each port arrival, it is updated according to:

$$T(port_t, state) \leftarrow (1 - \alpha_T) T(port_t, state) + \alpha_T m(state) \text{ for all states,}$$

using learning rate α_T and a memory trace vector, m, of the most recent path into the port, reflecting each hex traversed on the current trial, discounted by the experienced distance from the port. The memory trace, m, is itself initialized to zeros at the start of each trial, then learned over the trial by discounting and accumulation at each timestep t:

$$m(state) \leftarrow \gamma_{MB} m(state) \text{ for all states}$$

$$m(state_t) \leftarrow 1$$

In this way, T reflects a model-based expected eligibility trace for possible paths to the port, comprising both experiential eligibility from the just-completed path into the port (analogous to TD(1)), and counterfactual eligibility arising from a recency-weighted average over previous port entries.[58,72]

To assess the ability of each learning model to capture animal behavior, we computed the likelihood of every left vs right choice taken at each hex by each rat, using the value estimates provided by the same dual-component hex learner. We assumed the following softmax choice rule:

$$P(c_{hex} = c \in L, R) \propto \exp(\beta V(s_c) + b_{persistence} I(c, lastchoice(hex))$$

where β is an inverse temperature and $V(s_c)$ is the value of the hex that would be arrived at next given a left or right choice, under the learning model. To capture the rats' tendency to repeat their previous choice, we also included a term $b_{persistence}$ acting as a bias towards the choice made on the most recent visit to the same hex, where $I(s, lastchoice(hex))$ is a binary indicator which is one for the choice made previously, zero for the other. We limited our analysis to binary choices encountered by the rats – times when rats entered a three-way intersection and exited through one of the two hexes to the rats' right or left.

Hex-state value TD(λ) learner

We also considered an alternative model for learning hex-state values, based on TD(λ). This algorithm maintained an eligibility trace of recently visited hex-states to propagate updates backwards at each timestep. By optimizing the trace decay parameter, λ, to fit the observed DA at each timestep, we could estimate the spatial extent of value updates, on average. Value learning was implemented according to the following rules:

$$e(state) \leftarrow \lambda\gamma e(state) \text{ for all states}$$

$$\textit{At non-port hexes}: e(state_t) \leftarrow 1$$

$$\textit{Learning rule at each hex transition} : V(state) \leftarrow V(state) + \alpha e(state)\delta_t \text{ for all states}$$

$$\textit{with} : \delta_t = \gamma V(state_{t+1}) - V(state_t)$$

$$\textit{Learning rule upon port entry} : V(state) \leftarrow V(state) + \alpha e(state)\delta_t \text{ for all states}$$

$$\textit{with} : \delta_t = R_t - V(state_t)$$

$$\textit{Upon reaching port} : e(state) \leftarrow 0 \text{ for all states}$$

Dopamine regression

We combined each learning model with a linear regression observation function to model the dopamine timeseries, i.e. $DA_t = \beta_0 + \beta_V V(state_t) + \in_t$ with noise $\in_t \sim Normal(0, \sigma^2)$. Here, the parameter β_V captures any covariation between modeled value and the measured dopamine timeseries.

Model fitting

We optimized the free parameters of the learning algorithms by embedding each of them within a hierarchical model to allow parameters to vary from session-to-session. Session-level parameters were themselves modeled as arising from a distinct population-level Gaussian distribution over sessions for each rat. We estimated the model to obtain best fitting session- and population-level parameters to minimize the negative log likelihood of the data using an expectation-maximization algorithm with a Laplace approximation to the session-level marginal likelihoods in the M step.[73] For hypothesis testing on population-level parameters (β_V), we computed an estimate of the information matrix over the population-level parameters, taking account of the so-called ''missing information'' due to optimization in the E-step,[74] itself approximated using the Hessian of a single Newton-Raphson step. For the behavioral choice model, fitting was performed similarly to DA regression in order to maximize the likelihood of observed choices (using the same learning model as for DA, but re-estimating all free parameters to fit the choices).

For the value-iteration algorithm, which only sought to estimate the discount factor, γ, we used a simpler function-minimization protocol. On a session-by-session basis, we found the minimum of the negative log likelihood function of the DA data, given γ. As this was a simple scalar function, we used the minimize_scalar function from the SciPy package in Python. Parameter search was unbounded using Brent's algorithm, but γ values were rescaled between 0 and 1.

Model comparison

To isolate the contributions of each independent learning component, we created two nested models: one with α_{TD} and γ_{MF} both set to 0 (MB update only), and another with α_{MB}, α_T, and γ_{MB} all set to 0 (TD update only), and we compared each of these to the full model. In order to compare models with different numbers of free parameters, correcting for any bias due to overfitting, we computed a cross-validated approximation to the negative log marginal likelihood for each session.[36] Specifically, we used leave-one-session-out cross validation for the population-level prior parameters and a Laplace approximation for the per-session parameters: for each session, we refit the population-level model omitting that session, then conditional on that prior, we computed a Laplace approximation to that session's log marginal likelihood. We aggregated these per-session scores to obtain a total score for each rat and model. Finally, we use paired tests on these scores across rats, between models, to formally test whether any model fit consistently better over the population of rats. We depict relative fit subtracting out the dual-component model fit scores, so that positive values indicate superior dual-component model fit.

Data Analyses

Port-choice analyses

The frequencies of port visits and path choices were calculated using a five-trial rolling mean. To compute changes in visit frequency, we subtracted the mean frequency from the five trials prior to a block change from the frequencies after a block change. Note that paths here, and in most analyses, are defined by port visits (e.g., running from port A to port B), rather than specific sequences of hexes. ''Better'' and ''Worse'' ports were defined as those where the reward probability increased or decreased, respectively, compared to the prior block. This included changes from 10% reward probability to 50% reward probability, so the ''Better'' port was not necessarily the highest reward probability port in the maze. Similarly, ''Longer'' and ''Shorter'' paths were defined relative to the previous block, and paths whose length did not change were not included in this analysis.

All mixed-effects regression analyses were performed in R using the package lme4. Random effects were estimated over the levels of rat and session-within-rat. To identify any significant contributions of reward probability and path length on choice, we used a logistic mixed-effects regression of the following form:

$$\log(P(\text{choose left}) / (1 - P(\text{choose left})) = \beta_0 + \beta_1 \left(P(\text{reward})_{\text{left}} - P(\text{reward})_{\text{right}}\right) + \beta_2 * (\text{distance}_{\text{left}} - \text{distance}_{\text{right}}),$$

where the intercept captured any variation due to turn-direction bias. "Left" was defined on a trial-by-trial basis as the left of the two available ports, when oriented away from the previously visited port. For example, if the top port had just been visited, the bottom right port would be left, and the bottom left would be right. To avoid periods when rats are learning the probabilities of reward, we only included data from the second halves (> trial 25) of each block. Both probability differences and length differences were scaled between zero and one to compare effects in common units.

To isolate any effects of inference on port choice, we ran a similar logistic mixed-effects regression of port choice:

$$\log(P(\text{choose left})/(1 - P(\text{choose left})) = \beta_0 + \beta_1 * R_{t-1} + \beta_2 * R_{t-2} + \beta_3 * R_{t-3} + \beta_4 * R_{t-4} + \beta_5 * R_{t-5},$$

where "t-n" denotes prior trials where the left port was visited, and R denotes the reward outcome on that trial. Critically, we ran this regression for two subsets of data: trials where the rat took the same path to the goal port the last time it was visited, and trials where the rat last took an alternative path to the goal port. Paths, here, were defined based on the start and end ports, not the specific sequence of individual hexes traversed.

In addition, we sought to avoid possible confounds that arise due to decaying reward representations over time. For example, for a port that has not been visited in 10 trials, memory of the last outcome may have decayed, or uncertainty may have increased, compared to a port visited one trial ago (i.e., when a rat has been running back and forth between two ports and ignoring the other). To control for variations in the trial-lag length between traversals to the port of interest (the left option), we only included trials where the left available port was visited exactly two trials prior. This way, we are not comparing results from recent same-path reward to older alternative-path rewards, or vice versa.

Ramp analyses

Ramp slopes were estimated by fitting a linear regression model to the hex-level DA along the last 15 hexes traversed before port entry, in each session. The single rat that did not show significant positive ramping was not included in remaining analyses of DA ramping and value coding.

To scale and remove average ramps from individual-trial DA traces, we first calculated the average ramp over the last 10 hexes traversed for each rat. Because we were interested in scaling the entire ramp as a function of estimated gain, we needed to remove any negative values. To do this, we first rescaled each rat's average ramp between 0.1 and 0.9 (we refer to this as the control ramp, for clarity). For each path traversal of interest, we then fit a linear regression of the observed DA data to that rat's control ramp. An intercept captured remaining broad directional differences in the ramp (e.g., when the initial portion of the observed DA ramp was negative). We then scaled the control ramp by the estimated regression coefficient, added the intercept to the scaled control ramp, and subtracted this result from the DA trace. We were left with residual DA values, which we used for visualization in Figure 4L. To assess which portions of the observed propagating bump are significantly different than what can be expected by chance, we performed a permutation analysis. We computed null residuals along each path by shuffling the sequence of traversals, using equal numbers of traversals along the same paths as in the observed residual analysis. To estimate null distributions at each distance from the port, we computed 1000 shuffled null residual traces, and assessed the distribution at each distance (in hexes) from the destination port. Comparing observed residuals to the upper 95% confidence interval bounds allowed us to identify distances where residuals were significantly above chance.

Barrier-change dopamine analyses

To analyze discovery of a barrier change (either newly available or newly blocked) we aligned signals on first-detected entry into a hex immediately adjacent to the changed hex. At these hex transitions, the changed hex is readily visible. Initial new-hex exposures where the rat subsequently entered the new path were defined as those where the rat entered the newly available hex directly following its discovery.

DA regression analyses

We needed to isolate the hexes where values will differ depending on experience-based versus inference-based updates. To this end, we excluded all overlapping hexes between the same and alternative paths to the goal port. In other words, we only included the hexes prior to the first choice point on each trial where the rat has the opportunity to choose between the two available ports (see Figure 5).

To assess whether DA reflected the last reward outcome at the goal port following a traversal of the same path-dependent hexes and/or an alternative sequence of hexes, we ran a mixed-effects regression of the following form:

$$\text{Path-dependent DA} = \beta_0 + \beta_1 * R_{t-1} + \beta_2 * R_{t-2} + \beta_3 * R_{t-3} + \beta_4 * R_{t-4} + \beta_5 * R_{t-5},$$

where "t-n" denotes prior trials where the goal port was visited, and R denotes the reward outcome on that trial. Similar to the port-choice analysis, we ran this regression for two subsets of data: trials where the rat previously took the same path to the goal port, and trials where the rat took an alternative path to the goal port. Again, to control for biases that can arise due to differences in the number of trials since the port was last visited, we exclusively analyzed trials where the goal port was visited two trials ago. The inclusion of the prior five outcomes at the goal port controlled for DA scaling effects due to earlier rewards at the same port.

QUANTIFICATION AND STATISTICAL ANALYSES

Statistical tests and results are reported with any text introducing quantifications of results, both in the results section and in figure legends. Unless otherwise specified, we treated individual rats as the unit of analysis, rather than, e.g., fiber recording locations. Plots of aggregated data show mean +/- SEM, unless otherwise specified in the figure legends. Inclusion criteria for specific analyses are stated in both the STAR Methods and results sections. In general, rats were excluded from the dataset if their choice preferences did not significantly scale with expected reward or distance cost. Behavioral sessions were excluded if rats did not perform at least 100 trials. dLight fiber photometry recordings were excluded if a recording failure occurred at any point during the session, such as an optical fiber becoming broken or unplugged. We did not expect sufficient power to assess sex differences in this study, but we included both males (n=7) and females (n=3) in order to better identify findings that were robust across sexes. Rats were not assigned to separate experimental groups, so no blinding was performed. No sample size precalculation was performed.