

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Neural Solutions to the Credit Assignment Problem

Permalink

<https://escholarship.org/uc/item/9d5699m5>

Author

Krausz, Timothy Amos

Publication Date

2023

Peer reviewed|Thesis/dissertation

Neural Solutions to the Credit Assignment Problem

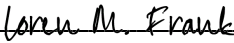
by
Timothy Amos Krausz

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in
Neuroscience

in the
GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:


DocuSigned by:

9C247AE9D5474C3... Loren M. Frank
Chair

DocuSigned by:

DocuSigned by:4F0... Joshua Berke

DocuSigned by:

DocuSigned by:33... Dr. Mazen Kheirbek

DocuSigned by:

12024A1A01564EB... Alexandra Nelson

Committee Members

Copyright 2023

by

Timothy Amos Krausz

Acknowledgements

While this thesis was by far the most solitary effort I've ever engaged in, I so often felt propped-up by the strong support network of many individuals who must be acknowledged here. First and foremost, I'd like to thank my scientific mentors. Among them, I am most grateful for my advisor, Dr. Josh Berke, who invested a truly commendable amount of time and effort in mentorship, including weekly meetings throughout the PhD. I will never forget the first months I spent in the lab, shooting design conceptions back and forth while scheming up our new behavioral task. He consistently encouraged me to think bigger, do better, and push myself outside of my comfort zone. While he held me to a high scientific standard, he also gave me the tools to meet it.

Josh was not the only busy scientist who dedicated many hours to intellectual discussions and mentorship. My PhD would not have been close to the same without Dr. Loren Frank and Dr. Nathaniel Daw. Thank you for all of the meetings (especially the many at odd hours during stints outside of the country), and for teaching me how to creatively approach questions with analytical rigor. Dr. Alexandra Nelson, Dr. Mazen Kheirbek, and Dr. Loren Frank, thank you for the time you've dedicated towards advice and guidance over the past five years as valued members of my committee. Dr. Ali Mohebi, there are many, many, things I could thank you for, but most of all thank you for continually being the best scientific role model I could ask for. Thank you for being an example with your unending enthusiasm for the scientific pursuit, and your ability to filter out the noise that can be so loud in academia.

Thank you to everyone in my life who was patient with me during long days and missed events so that I might attempt to get one more recording that may or may not work. I am forever

grateful to my parents, Ron and Susan, who instilled in me a deep confidence to pursue and indulge my curiosity. Thank you for being an honest bouncing board during our many calls on long walks home, and for always encouraging me to invest time in what's important to me. I've never felt anything but supported by you.

Finally, thank you to the best partner I could ask for through this PhD. Beethoven, I don't want to say I couldn't have done this without you, but I guarantee it was exponentially better with you by my side. I'm not superstitious, but I swear every time you sent a text with "good rat vibes" I got the best recordings of my PhD. So, thanks for that as well.

Contributions

Chapter 1 was written by Timothy A. Krausz

Chapter 2 was adapted from Timothy A. Krausz, Alison E. Comrie, Ari E. Kahn, Loren M. Frank, Nathaniel D. Daw, Joshua D. Berke. Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron* (2023).

<https://doi.org/10.1016/j.neuron.2023.07.017>.

Chapter 3 was written by Timothy A. Krausz. Additional contributions to this work were provided by Daniel Egert, Alison E. Comrie, Eric L. Denovellis, Loren M. Frank, Nathaniel D. Daw, and Joshua D. Berke.

Neural Solutions to the Credit Assignment Problem

Timothy A. Krausz

Abstract

To survive in the natural world, animals must learn to predict which actions and places will produce the resources necessary for life. When such rewarding outcomes are obtained, the brain must decide which subset of actions and places deserve credit, and to what extent. Due to the large and complex action space, and the instability of natural environments, adaptively updating predictions of future reward (values) poses a formidable computational challenge. Dopamine (DA) in the nucleus accumbens (NAc) is a neuromodulatory signal whose documented dynamics position it as an ideal candidate neural value signal. Meanwhile, neural representations of place – both actual and simulated – in the dorsal hippocampus (dHip) provide a candidate mechanism to adaptively assign credit to places distant from reward. In this body of dissertation work, I first develop a complex, yet tractable, spatial foraging task that emulates the credit assignment challenges posed by the natural world: numerous actions that do not directly result in reward, unstable paths to rewarding locations, and probabilistic reward outcomes. By recording from NAc DA in this task using dLight fiber photometry, I find that NAc DA robustly scales with spatial estimates of value (“place values”). Leveraging this relationship, I identify two key valuation algorithms used to generate this value signal: progressive propagation over space using directly experienced outcomes, and maze-wide updates using inference. Next, in a subset of rats, I simultaneously record from NAc DA using dLight fiber photometry and dHip pyramidal neurons using a custom 256-channel silicon electrophysiology probe. I achieve millisecond-timescale decoding of dHip place representations using a novel two-dimensional

spatial state-space algorithm. These included 8Hz “theta”-associated sweeps ahead of the animal into available paths, and representations of distant locations following reward. When dHip represented a higher-value available future path, NAc DA increased more than when dHip represented a lower-value path. As evidence for value updates, if a location was represented in dCA1 following reward, NAc DA was higher the next time the rat traversed that location, compared to traversals when that location had not previously been represented. These preliminary results provide striking new evidence for specific neural mechanisms that implement inference through simulation, and may therefore underlie intelligent learning and decision making. In all, this body of work provides novel insights into the neural mechanisms responsible for generating predictions of future reward to adaptively guide behavior.

Table of Contents

| | |
|---|-----|
| CHAPTER 1: INTRODUCTION | 1 |
| <i>REFERENCES</i> | 13 |
| CHAPTER 2: DUAL CREDIT ASSIGNMENT PROCESSES UNDERLIE DOPAMINE SIGNALS IN A COMPLEX SPATIAL ENVIRONMENT | 24 |
| <i>REFERENCES</i> | 71 |
| CHAPTER 3: EVIDENCE THAT DOPAMINE-ENCODED VALUES ARE RETRIEVED AND UPDATED THROUGH MENTAL SIMULATION | 82 |
| <i>REFERENCES</i> | 109 |

List of Figures

| | |
|--|-----------|
| FIGURE 2.1: ADAPTIVE BEHAVIOR IN THE SPATIAL FORAGING TASK..... | 28 |
| FIGURE 2.2: DOPAMINE PULSES AT REWARDS AND NOVEL PATH OPPORTUNITIES | 30 |
| FIGURE 2.3: DA RAMPS REFLECT DYNAMIC EXPECTATIONS OF UPCOMING REWARD..... | 32 |
| FIGURE 2.4: PROGRESSIVE PROPAGATION OF DA SIGNALS ACROSS SPACE | 36 |
| FIGURE 2.5: MODEL-BASED INFERENCE GLOBALLY UPDATES DA PLACE VALUES AND GUIDES CHOICES | 39 |
| FIGURE 2.6: A COMBINED LOCAL AND GLOBAL VALUE-UPDATE MODEL ACCOUNTS FOR HEX-LEVEL DA | 42 |
| SUPPLEMENTAL FIGURE 2.1: NAVIGATIONAL ADAPTATIONS TO MAZE CONFIGURATION CHANGES | 65 |
| SUPPLEMENTAL FIGURE 2.2: EXTENDED ANALYSIS OF DA PULSES..... | 66 |
| SUPPLEMENTAL FIGURE 2.3: INDIVIDUAL-ANIMAL RECORDING LOCATIONS AND DA GOAL-APPROACH RAMPS | 68 |
| SUPPLEMENTAL FIGURE 2.4: FURTHER COMPARISON OF TD ALGORITHMS TO DA SIGNALS | 69 |
| SUPPLEMENTAL FIGURE 2.5: INDIVIDUAL-SESSION DUAL-COMPONENT RL MODEL COMPARISON..... | 70 |
| FIGURE 3.1: EXPERIMENTAL APPROACH AND PLACE RECORDING IN A COMPLEX MAZE TASK | 86 |

FIGURE 3.2: THETA-BASED PROSPECTIVE SWEEPS ARE ASSOCIATED WITH DA
RELEASE PROPORTIONAL TO THE VALUES OF REPRESENTED LOCATIONS.....86

FIGURE 3.3: POST-REWARD REPRESENTATIONS FOLLOW DA REWARD
RESPONSES AND UPDATE PLACE VALUES.....93

List of Abbreviations

DA – Dopamine

dHip – Dorsal Hippocampus

HPD – Highest Posterior Density

LFP – Local Field Potential

MB – Model Based

MF – Model Free

NAc – Nucleus Accumbens

PFC – Prefrontal Cortex

RL – Reinforcement Learning

RPE – Reward Prediction Error

SPN – Spiny Projection Neuron

SWR – Sharp-wave Ripple

TD – Temporal Difference

VTA – Ventral Tegmental Area

Chapter 1:

Introduction

The credit assignment problem

The natural world abounds with opportunities to act and goals to pursue, yet only a sparse subset of these will result in the resources necessary for survival. To obtain these critical resources, animals must string together numerous actions in sequence, often traversing many places that lack reward. As a result, most actions taken and places encountered will be separated from any downstream rewards by vast gaps in both space and time. Meanwhile, animals contend with several potentially detrimental costs, including the energetic cost of movement and the opportunity cost of staying put. A key challenge, then, is selecting which of the many available actions are worth taking, and which goals are worth pursuing. Adaptive evaluative processes would benefit greatly from accurate predictions of future reward (“value”). Lacking perfect knowledge of the world, the brain is instead forced to leverage its limited experience to predict which places are worth going and what actions are worth taking. Thus, when the rare reward is finally obtained, the brain must answer a difficult question: which subset of the innumerable actions from the past deserves credit for producing this current rewarding outcome? This is termed the “credit assignment problem,” whose adaptive execution can dictate the survival of the organism.

Theoretical approaches and solutions

To begin to understand what processes the brain might use to perform credit assignment, we can leverage insights from decades of hypothesized dichotomies used to explain animal learning and decision-making. In other words, what can animal behavior tell us about how the brain evolved to predict future rewards from past experiences? If we examine popular theories in over the years, we find a set of proposed dichotomies: goal-directed versus habitual,¹⁻⁵ flexible versus inflexible,⁶ and allocentric versus egocentric,⁷ to name some of the most common. In goal-directed behaviors, animals decide that a specific goal is valuable enough to pursue, so they assign value to the upstream actions that will likely lead to that goal. Habitual behavior, in contrast, arises from learning the association between a stimulus, the animal's response, and the subsequent outcome. After repeated stimulus-responses are followed by desirable outcomes, stimuli will induce the animal to respond in the same fashion, even when the desired outcome no longer results. At this point, the action has a high value independent of the outcome being pursued, and behavior becomes habitual. Such inflexible behaviors often overlap with those classified as stereotyped, where a stimulus results in an animal producing a predictable action or action sequence. A flexible decision-making strategy, on the other hand, would allow the animal to explore new actions that may lead to better outcomes. Flexibility is especially responsive to novel observations – e.g., the omission of a rewarding outcome following an action that previously produced reward. Animals using a flexible strategy would be able to devalue that action. In navigational decision-making, animals may use what's termed an "allocentric" framework, considering their position with reference to the broader environmental context. Alternatively, they might rely on their own frame of reference ("egocentric") to select actions. When I enter my local grocery store, for example, I can consider where I am within the store,

where the desired product is, and determine what direction will move me towards that product. Alternatively, I could use an egocentric reference frame and remember that moving to the right is the best action from my currently observed location.

While the terminology used to describe different evaluative strategies has varied, much of this variability can be explained by considering the credit-assignment process involved. For the question of which actions/places deserve credit, one type of process assigns credit to the actions that directly produced the rewarding outcome. If left unchecked, this will give rise to habits and inflexible behavior, and promote an egocentric framework. A distinct strategy would permit credit assignment to additional alternative actions that *could have* led to reward. This inferential option would allow animals to more adaptively evaluate alternative routes to reward, permitting goal-directed, flexible, and allocentric behaviors. For the question of *how much* credit to assign to different actions, one extreme would only subtly change the values of preceding actions with each observed outcome. This strategy would produce values consistent with long-running average outcomes, but single adverse outcomes would fail to change behavior, resulting in inflexibility and habits. At the other extreme, a credit-assignment strategy might severely increase or decrease the values of prior actions according to the most recently observed reward outcome. Under this regime, decisions could more flexibly adapt behavior following the observation of novel adverse or rewarding outcomes. Such a switch between habitual and goal-directed behaviors has similarly been described by a trade-off in speed versus accuracy.⁸

It is no surprise, then, that a framework to unify these various descriptive dichotomies⁹ was found in the largest body of theoretical work on solutions to the credit-assignment problem: the machine-learning field of Reinforcement Learning (RL).¹⁰ Like biological agents, RL agents operate under the guiding principle of reward maximization. A key feature of the classic RL

framework is the discretization of environments into sets of distinct states, or observations about the environment. Each state is associated with its own set of actions for the agent to choose from. Actions can lead the agent from one state to another, but only a sparse set of states tend to produce reward. Another key feature of RL is the ability to associate values with states and actions, and use these values to guide decisions. An intuitive example can be found in the hallmark RL task structure: grid-world environments.¹⁰⁻¹² In these two-dimensional spatial tasks, agents must learn to select the directional actions from each state that will lead them from a starting state to a rewarding state. Decades of work has been dedicated to the development of algorithms that learn which actions and which states deserve credit for obtaining reward.

Broadly, RL algorithms can be grouped into two categories. One class of algorithm relies on direct experience alone (of actions, states, and outcomes) to update value estimates. This strategy poses a minimal burden on compute time and resources. All that must be stored and retrieved are cached value estimates for each experienced action and/or state. As an example, an agent will transition from one state to another, observe whether the value of the new state is better or worse than expected, and adjust the prior state's value accordingly. This specific value-learning algorithm is termed temporal difference (TD) learning.¹³ Because they do not rely on the learning and maintenance of an internal representation ("model") of the environment, they are termed "model-free" (MF). As demonstrated by the TD example, MF algorithms can efficiently deploy fast and simple learning rules for credit assignment. At the other extreme, the second class of algorithm is able to store and leverage an internal model of the environment to update state values. By understanding which actions lead to which states, these model-based (MB) algorithms can rely on both direct and simulated experience. For example, an MB agent initialized at a novel starting point in a grid-world can simulate traversals through the

environment until it finds a worthwhile option. It can use these simulations evaluate both actions and goals. Then, following reward acquisition, MB agents can assign credit to the states that directly led to the current state, in addition to states that *could have* led to the current state.

Through a distinction between MF and MB algorithms, RL offers a single framework to explain the various proposed behavioral dichotomies, while also providing a fruitful source for the generation of hypothesized neural credit-assignment algorithms. Consider the case where an agent has performed the same action sequence leading to a rewarding state over many repeated episodes. At this point, both MF and MB algorithms can appropriately assign higher values to the states along that sequence. MB algorithms might be superfluous or even costly in this situation, but what happens when a barrier blocks that path, or the location of the reward changes? Upon observing the altered environment, the MB algorithm will be able to flexibly revalue states and actions with respect to the goal of interest. It can update values so that the agent favors novel routes around barriers or even shortcuts through the environment. The MF agent, on the other hand, will have to directly experience alternate actions, states, and outcomes before it effectively revalues the relevant actions. In extreme cases, the previously high-valued action sequence may be so rigidly valued that the stereotyped behavior persists as if exhibiting a habit. Finally, by tuning a learning rate parameter, both class of algorithm are able to adjust the weight given to the most recently experienced reward compared to the long-running average. The field of RL, therefore, not only explains distinct learning strategies, but also provides a framework to model credit assignment algorithms so that we may better understand neural solutions to this formidable computational challenge.

Hypothesized neural implementation

Identifying a candidate neural value signal.

If we are to investigate credit assignment in the brain, we first need to select a candidate value signal. My aim is to first characterize a neural signal that predicts future reward, and then ask what algorithms are used to generate these predictions. The ideal candidate would not only scale with estimates of future reward, but it should also be found in a brain region associated with evaluative processes. For this, I turn to the Nucleus Accumbens (NAc). The NAc is situated in the ventral portion of the striatum, the input nucleus of the basal ganglia.¹⁴ The primary excitatory afferents to the NAc originate in the prefrontal cortex (PFC), hippocampus, and basolateral amygdala,¹⁵⁻¹⁸ creating a hub for cognitive, spatial, and affective information processing.¹⁹ The NAc also receives dense dopaminergic innervation from neurons originating in the ventral tegmental area (VTA).^{20,21} In fact, NAc spiny projection neurons (SPNs) are typically classified into two distinct groups depending on the type of dopamine (DA) receptor they express.²² In D1 receptor expressing SPNs, DA has a positive modulatory effect,²³⁻²⁵ facilitating neuron excitability as fast as 100s of milliseconds later.²⁶ In D2 receptor expressing SPNs, DA has a negative modulatory effect, tampering neuron excitability.²⁷ D1- and D2-type SPNs also generally display distinct projection targets.^{22,28} NAc DA, therefore, is poised to exert significant modulatory effects over neural firing and downstream behavior.

The NAc has been causally implicated in crucial evaluative decision processes, many of which are DA dependent. Lesions of the NAc in maze tasks resulted in impaired spatial navigation.^{29,30} Intriguingly, DA receptor antagonists produced a similar effect.³¹ Upon examination in more constrained behaviors, it was revealed that the NAc is required for flexibly motivating approach towards distant goals, where DA in the NAc is critical for this key

function.³² When a cue signals that a lever is available to interact with an operandum (the goal in this case), NAc DA receptor blockade inhibits both the NAc neural response³³ and the approach behavior itself.^{34,35} DA has also long been implicated in effort-related decision-making, where DA blockade does not obliterate goal pursuit altogether, but induces a preference for the low-reward and lesser-effort option.^{36–39} DA in the NAc, therefore, is a critical signal for the brain's ability to evaluate whether goals are worth pursuing.

If we investigate the NAc DA dynamics themselves, we find this neuromodulator is an ideal candidate to signal value. In both spatial⁴⁰ and operant tasks,^{41–44} NAc DA gradually increases in a ramp as rats navigate towards expected rewards. Such a ramping dynamic is noteworthy, as gradually increasing ramps can implement a value signal discounted over space. As mentioned above, animals must take into account the energetic cost of travel. Thus, the animal may want to devalue (or “discount”) those future rewards that are more distant, compared to those that are close by.⁴⁵ In RL theory, ramping value signals are commonly used to implement discounting, where value typically decays exponentially with distance from the reward state.^{10,42} The magnitude of ramping RL value signals also scales with the magnitude of the expected reward at the goal state. If an agent is approaching a more rewarding location, they should have a higher value estimate. In a simple maze task with distinct reward magnitudes at separate goal locations, ramping DA signals were also found to scale with expected reward magnitude at the goal.⁴⁰ Indeed, previous work from our lab found that NAc DA dynamics in an operant task can be explained by estimates of value.^{42,43} How, then, is this value signal generated?

To probe the underlying credit assignment algorithms used to generate value estimates in NAc DA, we first need to address existing task limitations that have prevented a thorough examination of the subject. As discussed, credit assignment in the natural world proves so

difficult due to the complexities of animal's environments. A multitude of places and actions do not directly result in reward, paths between rewarding locations are numerous and unstable, and the rewarding resources themselves are unstable. Emulating such complex features in a controlled environment is intimidating, and it may even appear intractable. For this reason, the vast majority of our knowledge of neural value signals described above comes from highly constrained, controllable, and simplified tasks. These have primarily included simplified maze tasks with largely stable reward contingencies,^{40,46} and operant tasks with dynamic reward contingencies but highly simplified environments.^{42,47,48} This has proven powerful for understanding fundamental principles of reward learning, but the tasks have proven limiting for understanding how credit is assigned over places and actions with each new experience. The ideal task to identify neural algorithms for credit assignment would emulate the complexities of the natural world, while providing a tractable framework to study how value estimates evolve over time. To leverage the powerful insights from RL theory, the task should also be amenable to computational modeling of decision processes with RL algorithms. A grid-world type structure, for example, would allow for a one-to-one mapping between RL variables and neural measurements.

A candidate neural substrate for model-based credit assignment.

Evidence for internal model usage in animals can be dated back to Edward Tolman in 1948,⁴⁹ and it has only grown since. Through careful experiments with a variety of maze tasks, Tolman found that rats are able to execute navigational decisions that rely on an understanding of their place in relation to the rest of the environment. He referred to this representation as a

“cognitive map.” Since then, there have been countless reports of cognitive map usage across species.⁵⁰

Brain lesion and manipulation studies have identified one region as particularly critical for cognitive map usage: the hippocampus. Early work in the field found deficits in spatial learning and memory following hippocampal lesions.⁵¹ One insightful study identified the hippocampus as necessary for learning to navigate through familiar environments towards hidden goals,⁵² a strategy that requires some knowledge of how different places are related in space. Indeed, the hippocampus has long been hypothesized as the substrate for cognitive map processing.⁵³ Further lesion studies identified the dorsal subregion of the rodent hippocampus (dHip) is especially important for this type of learning,^{54–56} permitting rats to favor an allocentric over egocentric navigation strategy when useful.⁵⁷ It is not surprising, then, that a clever optogenetic experiment identified the dHip as an essential node in MB decision-making processes.⁵⁸ This leaves us with the question, what role do neurons in the dHip play in MB credit assignment?

Neural representations in the dorsal hippocampus (dHip) provide a candidate neural substrate for the implementation MB simulations. As previously mentioned, one powerful function of MB strategies is their ability to simulate experiences of actions, paths, and states, without expending the energetic and opportunity cost to attempt the same experiences in reality. Such simulations can be used to look ahead through possible future routes in order to evaluate and select the most favorable option. They can also be used following reward to represent possible paths to the rewarding location, so that the animal is more prepared to select those paths in the future. To perform such simulations of possible states, the candidate neural substrate must be one whose neurons represent the states within an environment. In spatial tasks, dHip

pyramidal neurons selectively fire in distinct spatial receptive fields (“place fields”) as animals navigate through the environment.^{59–61} Ensembles of these “place cells” form a coherent representation of an animal’s actual location within an environment, sequentially firing as animals traverse through their respective place fields (“local” place coding).⁶² In effect, dHip pyramidal neurons signal where within the environment the animal currently is. They send this signal to a breadth of downstream areas for further information processing, including value-associated regions of the cortex⁶³ and the NAc.^{16,64} However, dHip pyramidal neurons will also fire when rats are outside of the associated place fields. This extra-field firing was found to coherently encode other possible locations within the same environment at compressed timescales.^{65–70} Consistent with simulation, these “non-local” representations will replay paths the rat has taken,^{69–73} or represent paths the rat could take in the future.^{74–79} Some of these even predict the path the rat is about to take, as if planning the route to a goal.⁷⁸ When rats approach decision points, the representation has been observed to sweep into the available paths,^{79,80} as if querying possible futures for evaluation.

While these phenomena are consistent with mental simulation, whether non-local representations are used to implement credit assignment across space remains a critical gap in our understanding. This gap is due, in large part, to a combination of recording and task limitations. Decoding location with high spatial precision from dHip neural activity requires the simultaneous recording of numerous pyramidal cells.^{78,79,81} Prior recording strategies have relied primarily on independently-movable tetrodes, which are quite effective but usually demand large amounts of hardware atop the subject’s head. While possible, simultaneous recordings with other brain regions have proven to be a difficult endeavor. Further, for the same reasons described

above, the spatial tasks used to study non-local place coding limit the experimenter's ability to study dynamic neural valuation processes.

Experimental Approach

In this body of dissertation research, I approached the neural credit-assignment problem through a combination of behavioral-paradigm development, neural recordings, and computational modeling. First, I devised a novel spatial foraging task for rats that both emulates the challenges posed to credit assignment in the natural world and mirrors the complex grid-worlds of RL theory (Chapter II). Using this task, I identify a scalar neural signal, dopamine in the nucleus accumbens, that scales with the values of spatial states.⁸² I then use a combination of analytical approaches, including RL modeling, to assess which algorithms the brain uses to generate the NAc DA value signal (Chapter II). Finally, I employ a novel multi-region and multi-modal recording approach in the same behavioral task to test a candidate neural substrate for model-based valuation: simulated place representations in the dorsal hippocampus.⁵³ Using this novel technological approach, I simultaneously record from both NAc DA and dHip neurons. In Chapter III, I describe experiments and preliminary results investigating whether dHip representations of simulated places^{66,69,70,78,79,83} are used to implement credit-assignment over space in the NAc DA value signal. I present evidence that simulations of possible futures during navigation may retrieve DA place values. I also show that, following reward receipt, simulations of distant places in the maze are used to assign credit to the DA value representation of those places. In all, this body of dissertation work provides the field with a more complete understanding of neural solutions to the credit-assignment problem. Elucidating the mechanisms underlying this fundamental function is a critical step towards better understanding the fine line

between adaptive and maladaptive reward predictions, such as addictive or compulsive behaviors.

References

1. Rescorla, R.A. (1987). A Pavlovian analysis of goal-directed behavior. *Am. Psychol.* 42, 119–129.
2. Frese, M., and Sabini, J. (2021). *Goal Directed Behavior: The Concept of Action in Psychology* (Routledge).
3. Wood, W., and Rünger, D. (2016). Psychology of Habit. *Annu. Rev. Psychol.* 67, 289–314.
4. Marx, M.H. (1950). A stimulus-response analysis of the hoarding habit in the rat. *Psychol. Rev.* 57, 80–93.
5. Dezfouli, A., and Balleine, B.W. (2013). Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Comput. Biol.* 9. 10.1371/journal.pcbi.1003364.
6. Sylvie Granon, S.F. (2009). Functional neuroanatomy of flexible behaviors in mice and rats. In *Endophenotypes of Psychiatric and Neurodegenerative Disorders in Rodent Models*.
7. Klatzky, R.L. (1998). Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections. In *Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, C. Freksa, C. Habel, and K. F. Wender, eds. (Springer Berlin Heidelberg), pp. 1–17.
8. Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* 7. 10.1371/journal.pcbi.1002055.

9. 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.
10. Sutton, R.S., and Barto, A.G. (2018). *Reinforcement Learning, second edition: An Introduction* (MIT Press).
11. Onda, H., and Ozawa, S. (2009). A reinforcement learning model using macro-actions in multi-task grid-world problems. In *2009 IEEE International Conference on Systems, Man and Cybernetics (IEEE)*. 10.1109/icsmc.2009.5346139.
12. Bamford, C., Jiang, M., Samvelyan, M., and Rocktäschel, T. (2022). GriddlyJS: A Web IDE for Reinforcement Learning. *arXiv [cs.AI]*, 15051–15065.
13. Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.* 3, 9–44.
14. Dudman, J.T., and Gerfen, C.R. (2015). *The Basal Ganglia*.
15. Berendse, H.W., and Groenewegen, H.J. (1990). Organization of the thalamostriatal projections in the rat, with special emphasis on the ventral striatum. *J. Comp. Neurol.* 299, 187–228.
16. Groenewegen, H.J., Van der Zee, E.V., te Kortschot, A., and Witter, M.P. (1987). Organization of the projections from the subiculum to the ventral striatum in the rat. A study using anterograde transport of Phaseolus vulgaris leucoagglutinin. *Neuroscience* 23, 103–120.

17. Groenewegen, H.J., Room, P., Witter, M.P., and Lohman, A.H.M. (1982). Cortical afferents of the nucleus accumbens in the cat, studied with anterograde and retrograde transport techniques. *Neuroscience* 7, 977–996.
18. Groenewegen, H.J., Wright, C.I., and Beijer, A.V.J. The nucleus accumbens: gateway for limbic structures to reach the motor system? *Prog. Brain Res.* 107.
19. Floresco, S.B. (2015). The Nucleus Accumbens: An Interface Between Cognition, Emotion, and Action. *Annu. Rev. Psychol.* 10.1146/annurev-psych-010213-115159.
20. Ikemoto, S. (2007). Dopamine reward circuitry: Two projection systems from the ventral midbrain to the nucleus accumbens–olfactory tubercle complex. *Brain Res. Rev.* 56, 27–78.
21. Breton, J.M., Charbit, A.R., Snyder, B.J., Fong, P.T.K., Dias, E.V., Himmels, P., Lock, H., and Margolis, E.B. (2019). Relative contributions and mapping of ventral tegmental area dopamine and GABA neurons by projection target in the rat. *J. Comp. Neurol.* 527, 916–941.
22. Gerfen, C.R., and Surmeier, D.J. (2011). Modulation of striatal projection systems by dopamine. *Annu. Rev. Neurosci.* 34, 441–466.
23. Ericsson, J., Stephenson-Jones, M., Pérez-Fernández, J., Robertson, B., Silberberg, G., and Grillner, S. (2013). Dopamine differentially modulates the excitability of striatal neurons of the direct and indirect pathways in lamprey. *J. Neurosci.* 33, 8045–8054.
24. Ding, L., and Perkel, D.J. (2002). Dopamine modulates excitability of spiny neurons in the avian basal ganglia. *J. Neurosci.* 22, 5210–5218.

25. Nicola, S.M., Surmeier, J., and Malenka, R.C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annu. Rev. Neurosci.* *23*, 185–215.
26. Lahiri, A.K., and Bevan, M.D. (2020). Dopaminergic Transmission Rapidly and Persistently Enhances Excitability of D1 Receptor-Expressing Striatal Projection Neurons. *Neuron* *0*, 1–14.
27. Hernández-López, S., Tkatch, T., Perez-Garci, E., Galarraga, E., Bargas, J., Hamm, H., and James Surmeier, D. (2000). D2 Dopamine Receptors in Striatal Medium Spiny Neurons Reduce L-Type Ca²⁺ Currents and Excitability via a Novel PLCβ1–IP3–Calcineurin-Signaling Cascade. *J. Neurosci.* *20*, 8987–8995.
28. Pettibone, J.R., Yu, J.Y., Derman, R.C., Faust, T.W., Hughes, E.D., Filipiak, W.E., Saunders, T.L., Ferrario, C.R., and Berke, J.D. (2019). Knock-in rat lines with cre recombinase at the dopamine d1 and adenosine 2a receptor loci. *eNeuro* *6*. 10.1523/ENEURO.0163-19.2019.
29. Seamans, J.K., and Phillips, A.G. (1994). Selective memory impairments produced by transient lidocaine-induced lesions of the nucleus accumbens in rats. *Behav. Neurosci.* *108*, 456–468.
30. Floresco, S.B., Seamans, J.K., and Phillips, A.G. (1997). Selective Roles for Hippocampal, Prefrontal Cortical, and Ventral Striatal Circuits in Radial-Arm Maze Tasks With or Without a Delay.
31. Ploeger, G.E., Spruijt, B.M., and Cools, A.R. (1994). Spatial localization in the Morris water maze in rats: Acquisition is affected by intra-accumbens injections of the

- dopaminergic antagonist haloperidol. *Behavioral Neuroscience* 108, 927–934.
10.1037/0735-7044.108.5.927.
32. 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. *Journal of Neuroscience*. 10.1523/JNEUROSCI.3958-10.2010.
 33. Hoffmann, J.D., and Nicola, S.M. (2014). Dopamine Invigorates Reward Seeking by Promoting Cue-Evoked Excitation in the Nucleus Accumbens. *J. Neurosci.*
10.1523/JNEUROSCI.3492-14.2014.
 34. Nicola, S.M., Taha, S.A., Kim, S.W., and Fields, H.L. (2005). Nucleus accumbens dopamine release is necessary and sufficient to promote the behavioral response to reward-predictive cues. *Neuroscience*. 10.1016/j.neuroscience.2005.06.088.
 35. Yun, I.A., Nicola, S.M., and Fields, H.L. (2004). Contrasting effects of dopamine and glutamate receptor antagonist injection in the nucleus accumbens suggest a neural mechanism underlying cue-evoked goal-directed behavior. *Eur. J. Neurosci.* 20, 249–263.
 36. 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.* 10.1016/0166-4328(95)00151-4.
 37. Salamone, J. (1999). Interference with accumbens dopamine transmission makes rats more sensitive to work requirements but does not impair primary food reinforcement. *Behav. Pharmacol.* 10, S79.

38. 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.
39. Aberman, J.E., and Salamone, J.D. (1999). Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience* 92, 545–552.
40. 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*. 10.1038/nature12475.
41. 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, 1–15.
42. 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.
43. 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.
44. 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.

45. Healt, G. (2007). Discounting: A Review of the Basic Economics. The University of Chicago Law Review.
46. 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.
47. 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.
48. 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.
49. Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208.
50. Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* 100, 490–509.
51. Olton, D.S., Walker, J.A., and Gage, F.H. (1978). Hippocampal connections and spatial discrimination. *Brain Res.* 10.1016/0006-8993(78)90930-7.
52. Morris, R.G.M., Garrud, P., Rawlins, J.N.P., and O’keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions (McGraw-Hill).
53. O’Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*.

54. Moser, E., Moser, M.-B., and Andersen, P. (1993). Spatial Learning Impairment Parallels the Magnitude of Dorsal Hippocampal Lesions, but Is Hardly Present following Ventral Lesions.
55. Hock, B.J., Jr, and Bunsey, M.D. (1998). Differential effects of dorsal and ventral hippocampal lesions. *J. Neurosci.* *18*, 7027–7032.
56. Zhang, W.-N., Pothuizen, H.H.J., Feldon, J., and Rawlins, J.N.P. (2004). Dissociation of function within the hippocampus: effects of dorsal, ventral and complete excitotoxic hippocampal lesions on spatial navigation. *Neuroscience* *127*, 289–300.
57. Rogers, J.L., and Kesner, R.P. (2006). Lesions of the dorsal hippocampus or parietal cortex differentially affect spatial information processing. *Behav. Neurosci.* *120*, 852–860.
58. Miller, K.J., Botvinick, M.M., and Brody, C.D. (2017). Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* 10.1038/nn.4613.
59. O'keefe, J. (1976). Place Units in the Hippocampus of the Freely Moving Rat.
60. O'keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.*, 171–175.
61. Liberti, W.A., Schmid, T.A., Forli, A., Snyder, M., and Yartsev, M.M. (2022). A stable hippocampal code in freely flying bats. *Nature.* 10.1038/s41586-022-04560-0.
62. Wilson, M.A., and McNaughton, B.L. (1993). Dynamics of the Hippocampal Ensemble Code for Space.

63. Thierry, A.M., Gioanni, Y., Dégénétais, E., and Glowinski, J. (2000). Hippocampo-prefrontal cortex pathway: Anatomical and electrophysiological characteristics. *Hippocampus*. 10.1002/1098-1063(2000)10:4<411::AID-HIPO7>3.0.CO;2-A.
64. Trouche, S., Koren, V., Doig, N.M., Ellender, T.J., El-Gaby, M., Lopes-dos-Santos, V., Reeve, H.M., Perestenko, P.V., Garas, F.N., Magill, P.J., et al. (2019). A Hippocampus-Accumbens Tripartite Neuronal Motif Guides Appetitive Memory in Space. *Cell*. 10.1016/j.cell.2018.12.037.
65. Skaggs, W.E., McNaughton, B.L., Wilson, M.A., and Barnes, C.A. (1996). Theta Phase Precession in Hippocampal Neuronal Populations and the Compression of Temporal Sequences. *Hippocampus* 6, 149–172.
66. Foster, D.J., and Wilson, M.A. (2007). Hippocampal Theta Sequences. *Hippocampus* 17, 1–3.
67. Dragoi, G., and Buzsáki, G. (2006). Temporal Encoding of Place Sequences by Hippocampal Cell Assemblies. *Neuron* 50, 145–157.
68. Schmidt, R., Diba, K., Leibold, C., Schmitz, D., Buzsáki, G., and Kempfer, R. (2009). Single-Trial Phase Precession in the Hippocampus. 10.1523/JNEUROSCI.2270-09.2009.
69. Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440. 10.1038/nature04587.
70. Diba, K., and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* 10. 10.1038/nn1961.

71. Joo, H.R., and Frank, L.M. The hippocampal sharp wave-ripple in memory retrieval for immediate use and consolidation Nature reviews | NeurosciNce. Nat. Rev. Neurosci. 10.1038/s41583-018-0077-1.
72. Karlsson, M.P., and Frank, L.M. (2009). Awake replay of remote experiences in the hippocampus. Nature Publishing Group *12*. 10.1038/nn.2344.
73. Davidson, T.J., Kloosterman, F., and Wilson, M.A. (2009). Hippocampal replay of extended experience. Neuron *63*, 497–507.
74. 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.
75. 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. U. S. A. *117*, 689–697.
76. Gupta, A.S., van der Meer, M.A.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal replay is not a simple function of experience. Neuron *65*, 695–705.
77. Freyja Ólafsdóttir, H., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. Elife *4*, 1–17.

78. Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79.
79. 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.
80. Johnson, A., and Redish, A.D. (2007). Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *J. Neurosci.* 10.1523/JNEUROSCI.3761-07.2007.
81. Denovellis, E.L., Gillespie, A.K., Coulter, M.E., Sosa, M., Chung, J.E., Eden, U.T., and Frank, L.M. (2021). Hippocampal replay of experience at real-world speeds. *Elife* 10. 10.7554/eLife.64505.
82. Krausz, T.A., Comrie, A.E., Kahn, A.E., Frank, L.M., Daw, N.D., and Berke, J.D. (2023). Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron*. 10.1016/j.neuron.2023.07.017.
83. 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.

Chapter 2:

Dual Credit Assignment Processes Underlie Dopamine Signals in a Complex Spatial Environment

Introduction

Animals frequently make motivated choices based on prior experiences - for example, selecting paths towards 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. Temporal difference 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, or 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. ^{17,23,24}), thus avoiding the challenging credit-assignment question. In other paradigms, applying RL ideas involves unsupported arbitrary assumptions²⁵ – 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 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 (**Fig. 2.1**) 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 – 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 (**Fig. 2.1B**), and was closer (**Fig.**

2.1C), compared to the alternative. A mixed-effects logistic multiple regression, incorporating any turn biases (see 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 (**Fig. 2.1D**). After each block of 50-70 trials (traversals between ports), either the reward probabilities changed (**Fig. 2.1E**) or a barrier was moved to change available paths (**Fig. 2.1F**). After a change in reward probabilities, rats increased their choice of ports whose reward probability had increased (**Fig. 2.1G**). Following a barrier move, rats adjusted their port choices to favor shorter paths (**Fig. 2.1H**) and also progressively refined their specific paths to be more efficient (**Supp. Fig 2.1**).

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, dLight1.3b³² ($n = 10$ rats, 19 fiber locations, 82 behavioral sessions, 296 blocks, 16,379 trials, mean of 1638 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 (**Fig. 2.2B**). The magnitude of these phasic changes depended on port reward probability, in a direction consistent with RPE coding (**Fig. 2.2C**, Pearson correlation, rewarded trials mean coefficient = -0.221 ± 0.098 STDEV; omission trials mean coefficient = -0.111 ± 0.062 STDEV; 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 Methods). DA following port entry significantly scaled with

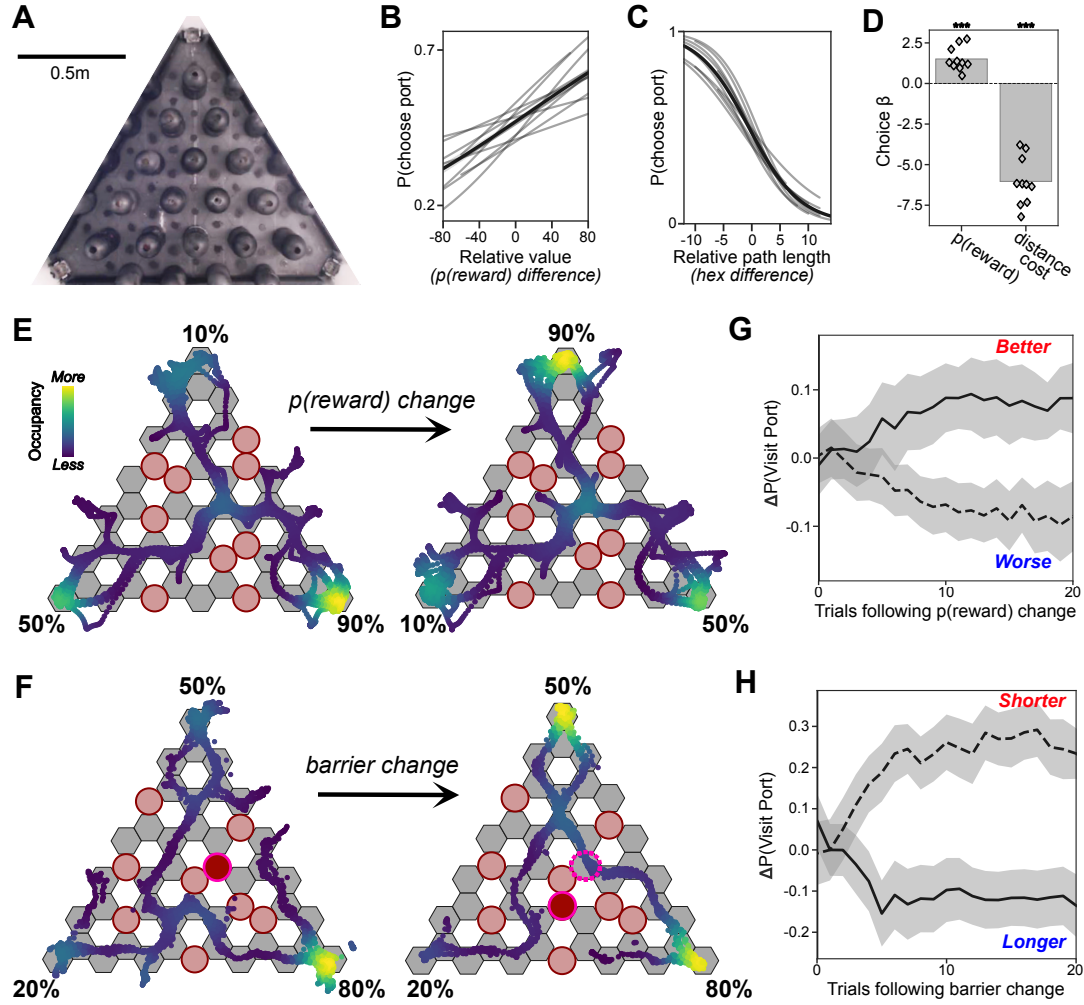


Figure 2.1. Adaptive behavior in the spatial foraging task.

(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 \pm 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 \pm SEM). “Trials” in (G) and (H) include only those where the rat had the opportunity to choose the port in question.

these RPE estimates (**Supp. Fig. 2.2**), although encoding of positive RPEs was notably stronger and more consistent across rats, compared to 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 no longer (**Fig. 2.2D-F**). 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 (**Fig. 2.2F**). 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 (**Supp. Fig. 2.2C**), and when the newly available hex was closer to the final destination port (**Supp. Fig. 2.2D**). However, these latter observations would require a larger data set 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 (**Fig. 2.3A**). This overall ramp was significantly positive in nine of ten individual animals (16/19 individual fibers; **Supp. Fig.**

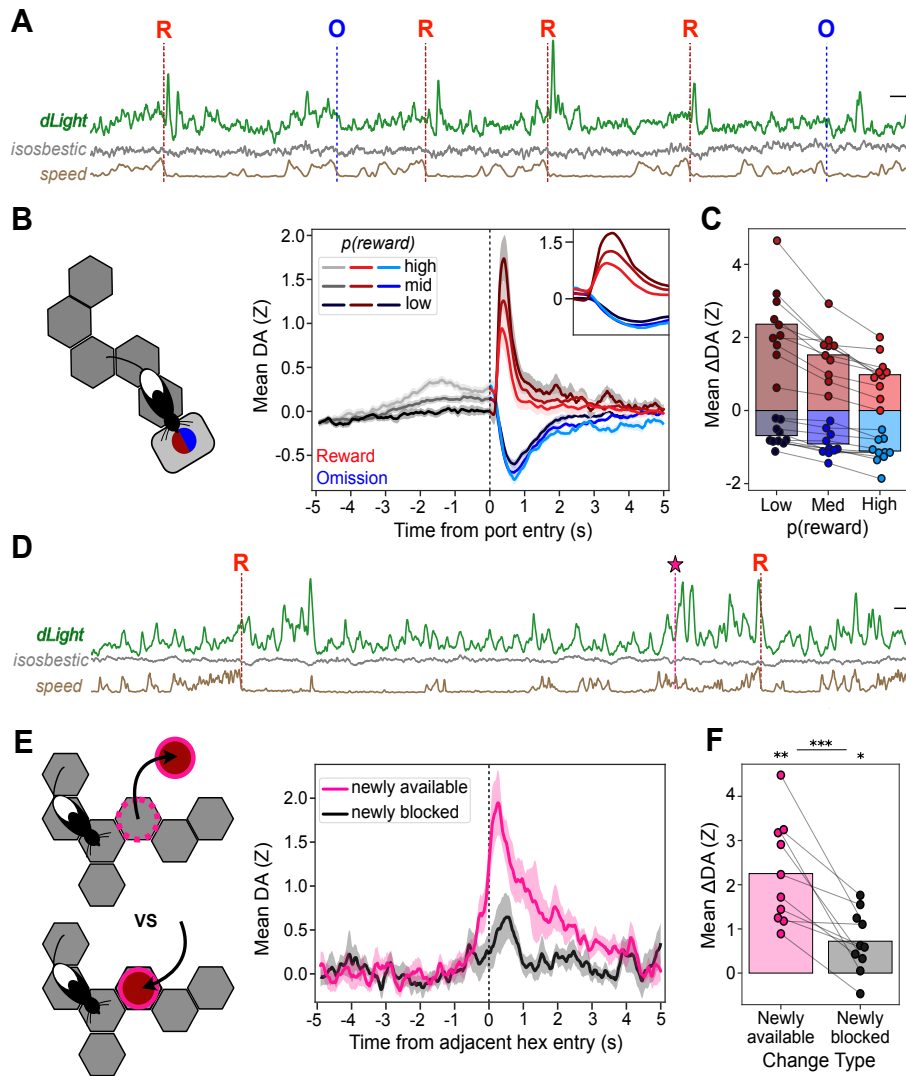


Figure 2.2. Dopamine pulses at rewards and novel path opportunities.

(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 \pm 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 \pm 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).

2.3A). To better understand the computations that give rise to this ramp, subsequent analyses

focused on those nine rats. The magnitude of the DA ramp scaled with the current reward probability of the approached port (**Fig. 2.3B**), 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 (**Fig. 2.3C**). 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 **Fig. 2.3C**), 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 (**Fig. 2.3D**). 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 towards higher-probability ports (**Fig. 2.3B**). However, running speed peaked later than DA (**Fig. 2.3A/B**), and cross-correlograms suggested that DA was predictive of speed (potentially driving the vigor of running) rather than merely reflecting it (**Supp. Fig. 2.3B**).

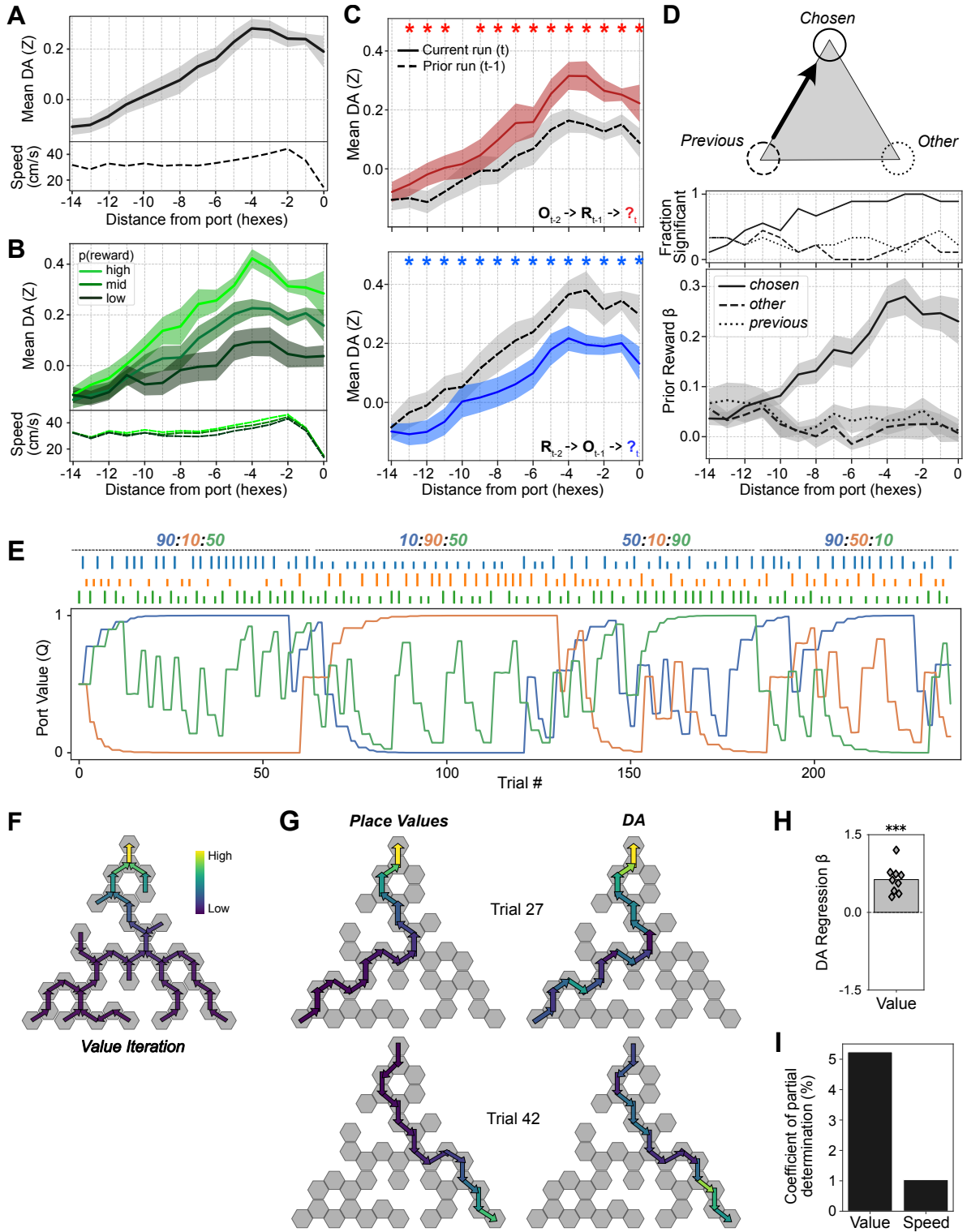


Figure 2.3. DA ramps reflect dynamic expectations of upcoming reward.

(Figure caption continued on the next page).

(Figure caption continued from the previous page). (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 $p(\text{reward})$ 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 \pm 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 *** $p < 0.001$.

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 (**Fig. 2.3E**), but we then distributed these values, discounted by spatial distance, throughout the maze ("value iteration";^{1,35} **Fig. 2.3F**; see Methods). The resulting hex-level pattern of value closely resembled DA on each trial (**Fig. 2.3G**), and a mixed-effects multiple regression analysis revealed a highly significant relationship between DA and these hex values (**Fig. 2.3H**, $p < 0.0001$, Likelihood Ratio Test, chi-square distribution with 77 degrees of freedom to account for each session-optimized γ value; see

Methods). This regression analysis also included running speed, yet hex values accounted for much more of the explained variability in the DA signal (**Fig. 2.3I**).

Over repeated trials, DA signals propagate backwards 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 tell-tale 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 state¹ (**Fig. 2.4A**). 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 backwards 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 (**Fig. 2.4B**). In simulations (see Methods), with TD(0) the reward evokes a value bump that propagates backwards over the course of multiple traversals (**Fig. 2.4C**). 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 (**Fig. 2.4D**). 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 (**Fig 2.4E,F**). 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 (**Fig. 2.4E**), in stark contrast to TD(1) (**Fig. 2.4F**).

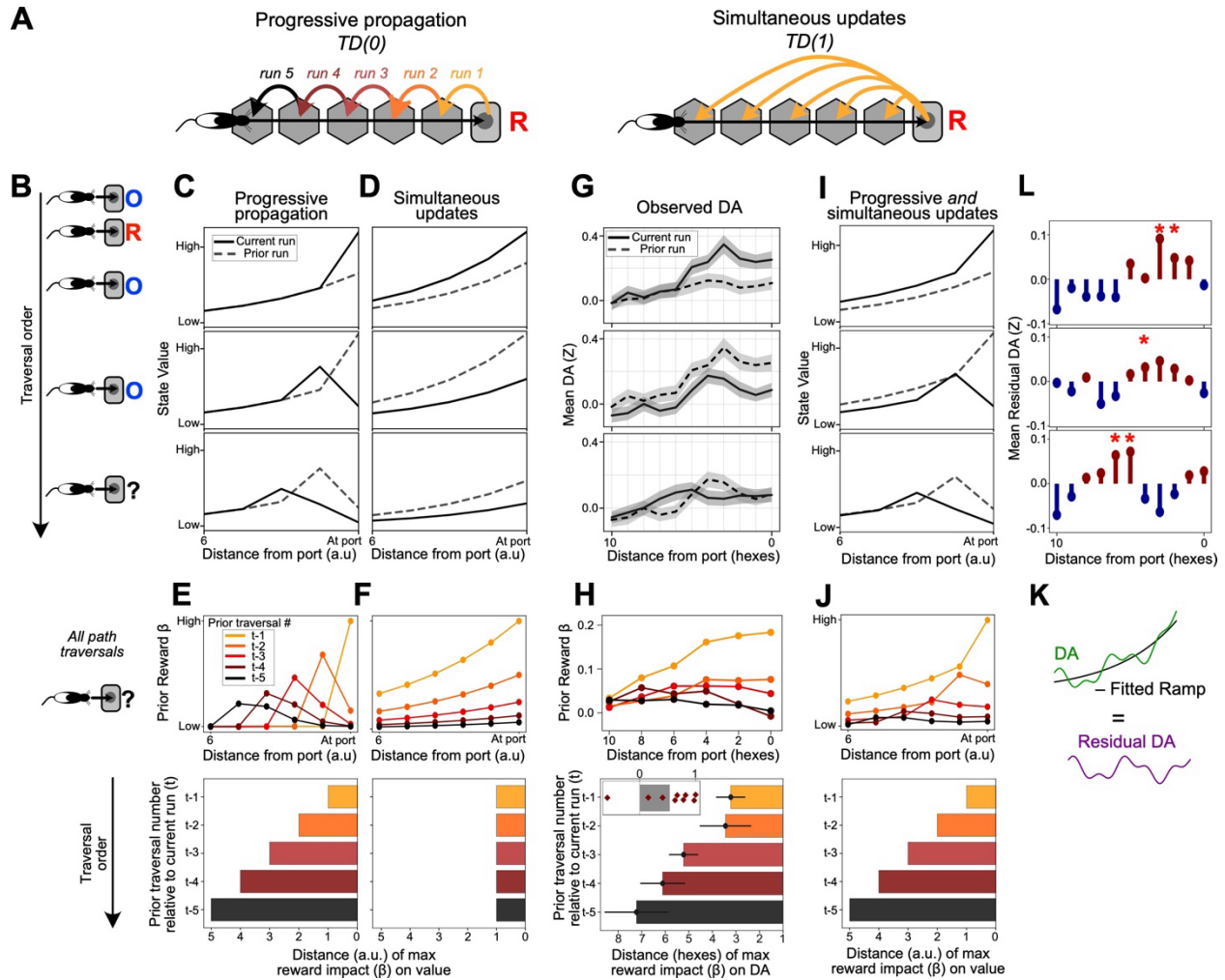


Figure 2.4. Progressive propagation of DA signals across space.

(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 \pm 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 \pm 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). (Figure caption continued on the next page).

(Figure caption continued from the previous page). (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 Methods).

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 **Fig. 2.4B**). The reward appeared to cause a spatial bump in DA, that moved further back from the reward port over successive traversals (**Fig. 2.4G**) – i.e., the key signature of TD learning with low λ . Next, we performed the multiple regression with all trial sequences, as in **Fig. 2.4E/F**, 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 (**Fig. 2.4H**; 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.

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 (**Fig. 2.3C**). Although each of these can arise separately as the extreme cases of TD(λ) (i.e. λ close to zero or one respectively), there is no intermediate setting of λ at which both of these patterns co-occur. Consistent with this, fitting a TD(λ) hex-state RL algorithm (see 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 (**Supp. Fig. 2.4A-E**). Fit λ numbers were consistently high for individual sessions (**Supp. Fig. 2.4B**), 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 (**Fig. 2.4I**). 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 (**Fig. 2.4J**).

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 Methods). As expected, removing the overall ramp left a residual DA signal that propagated backwards along the path over trials (**Fig. 2.4K/L**). These results are consistent with updating of DA value signals updated by at least two mechanisms – a TD(0)-like process responsible for backwards 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 (**Fig. 2.5A**). 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 (**Fig. 2.5B**). 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

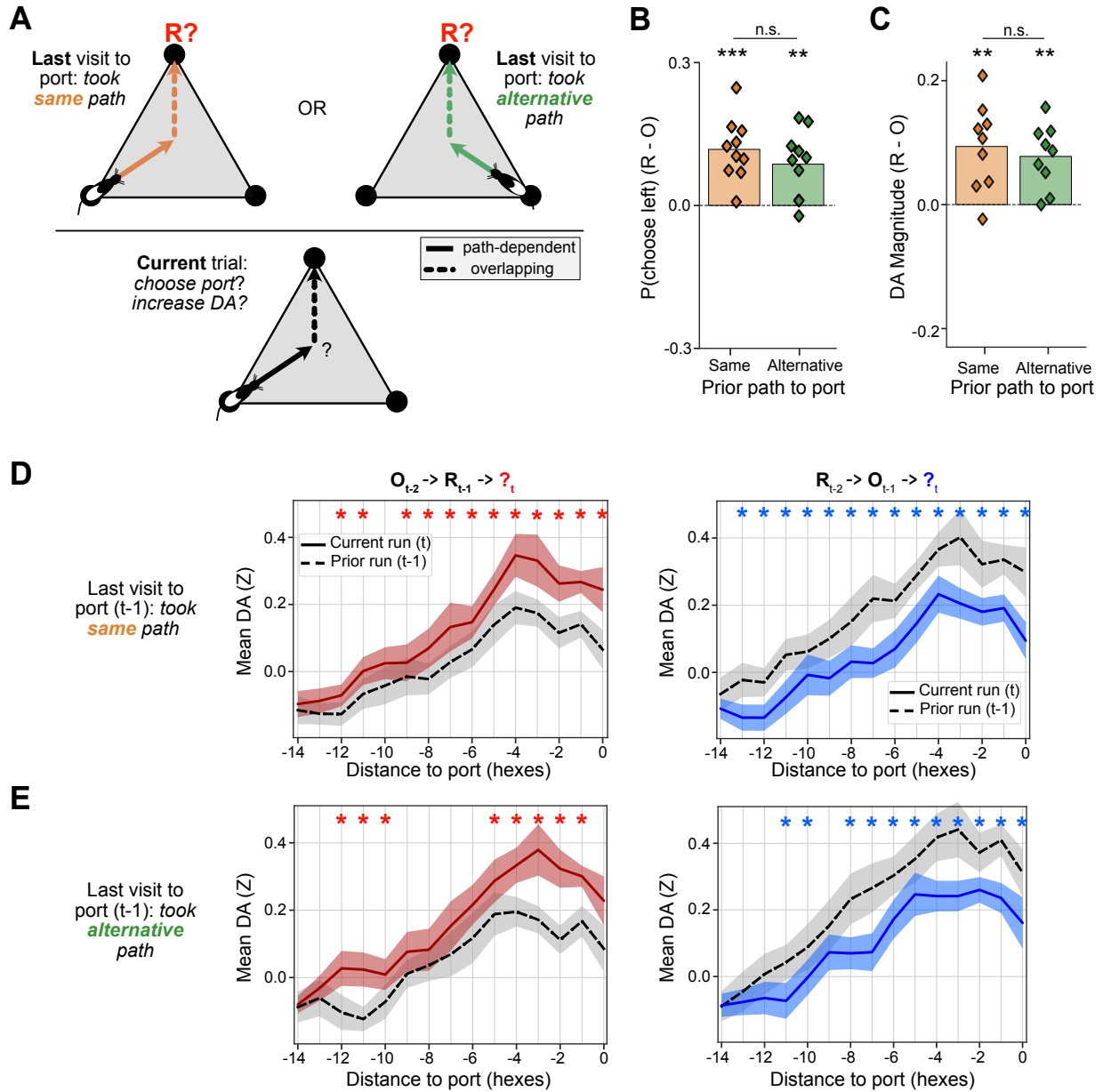


Figure 2.5. Model-based inference globally updates DA place values and guides choices.

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. (Figure caption continued on the next page).

(Figure caption continued from the previous page). 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).

five reward outcomes as features (see 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 model-based (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 (**Fig. 2.5A**). 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 (**Fig. 2.5C**). 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 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 reward had been obtained taking the same path, or the alternative (**Fig. 2.5D/E**). 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 (**Fig. 2.6**). 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; **Fig. 2.6A**; see Methods). Regression analysis revealed a significant relationship between values in this dual-process model and observed DA (mean $\beta = 0.798 \pm$

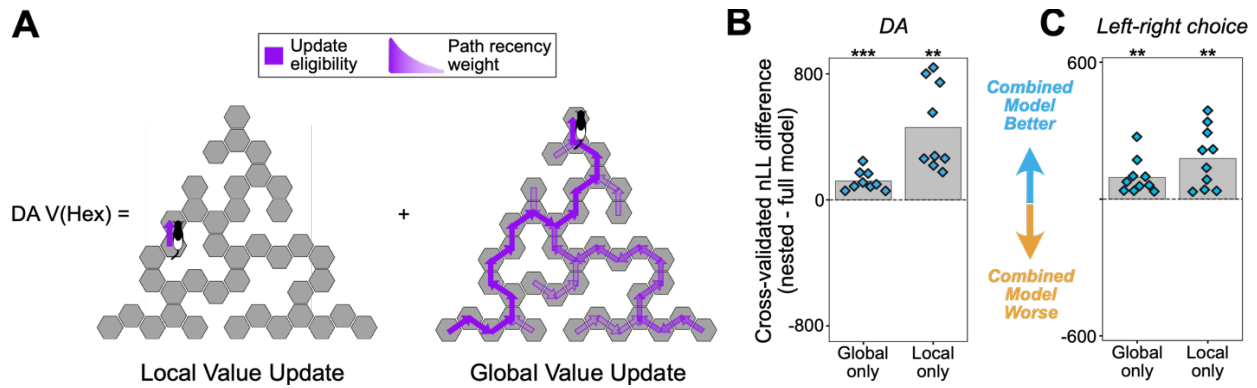


Figure 2.6. A combined local and global value-update model accounts for hex-level DA.

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).

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 to two nested models,³⁶ each with one update process removed (by setting the learning rate to zero). The combined model outperformed either process alone (**Fig. 2.6B**). This result was observed consistently across individual rats and sessions (**Supp. Fig. 2.5**), 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 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 (**Fig. 2.6C**).

Discussion

Theoretical models of reinforcement learning 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-maze¹⁹). 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 representations⁴⁰). 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 data set 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, here we found 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 afterwards, 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 signals⁴⁷ and may involve distinct time scales 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 make 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 assignment⁵¹). Evidence that DA ramping can be mechanistically distinct from RPE coding comes from a prior study in which we compared DA cell firing to release.⁵ Discrete reward cues evoked RPE-encoding burst firing of identified 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 (**Fig. 2.4**).

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 (McNamara et al. 2014) 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 towards goals (similar to another family of AI algorithms that use models for planning⁶²). 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 cycle⁶⁴ 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 (**Fig. 2.3**). 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 can reappear immediately if task contingencies change.¹⁸ There is analogous evidence that non-local activity in hippocampus declines over repeated experience, including both SWRs⁵⁴ 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

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 15 μ L 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 is visited (this rule is present throughout training and testing). Rats were then trained on the task until reaching criterion (\geq 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.⁷⁰ 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 2×10^{12} – was delivered using a stereotaxic injection pump (Nanject III). Virus was injected 200µm ventral to the target coordinates, as described in.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 reweighted Penalized Least Squares (airPLS) algorithm.⁷¹ 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.

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:

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

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:

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

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 function³⁶).

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 traces¹). 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-port states: $e(\text{state}_t) \leftarrow 1$

$$\delta_t = R_t + \gamma V(\text{state}_{t+1}) - V(\text{state}_t)$$

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

Upon reaching port: $e(\text{state}) \leftarrow 0$ 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 **Fig. 2.4**, 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 (**Fig. 2.4I/J**), 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:

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

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

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:

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

where α_{MB} is the model-based update learning rate, and $T(\textit{port}, \textit{state})$ 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:

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

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 :

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

$$m(\text{state}_t) \leftarrow 1$$

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:

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

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:

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

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

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

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

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

$$\text{with: } \delta_t = R_t - V(\text{state}_t)$$

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

Dopamine regression: We combined each learning model with a linear regression observation function to model the dopamine timeseries, i.e. $DA_t = \beta_0 + \beta_V V(\text{state}_t) + \epsilon_t$ with noise $\epsilon_t \sim \text{Normal}(0, \sigma^2)$. 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.

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:

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

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 **Fig. 2.4L**. 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 **Fig. 2.5**).

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:

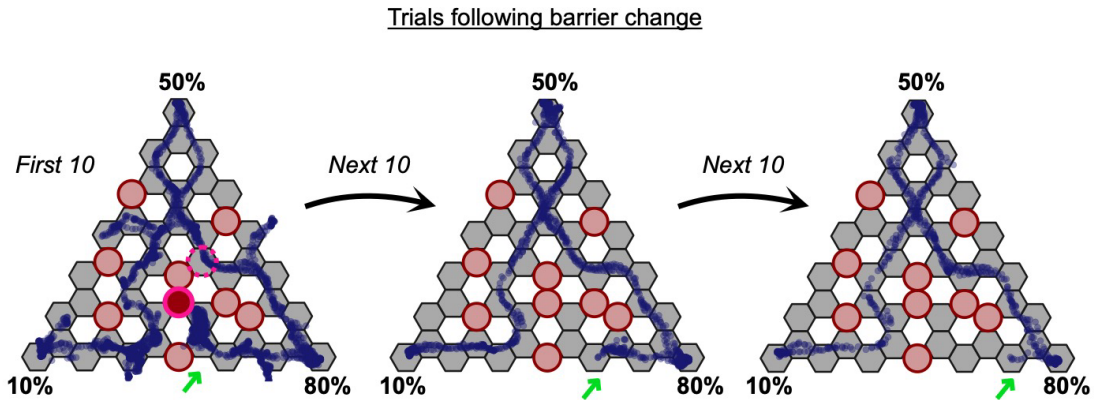
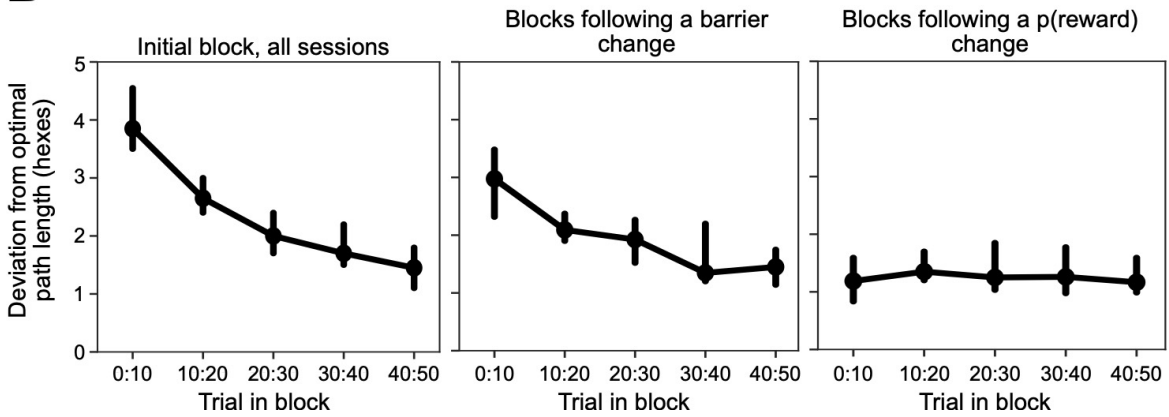
$$\text{Path-dependent DA} = \beta_0 + \beta_1 * R_{t-1} + \beta_2 * R_{t-2} + \beta_3 * R_{t-3} + \beta_4 * R_{t-4} + \beta_5 * R_{t-5},$$

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.

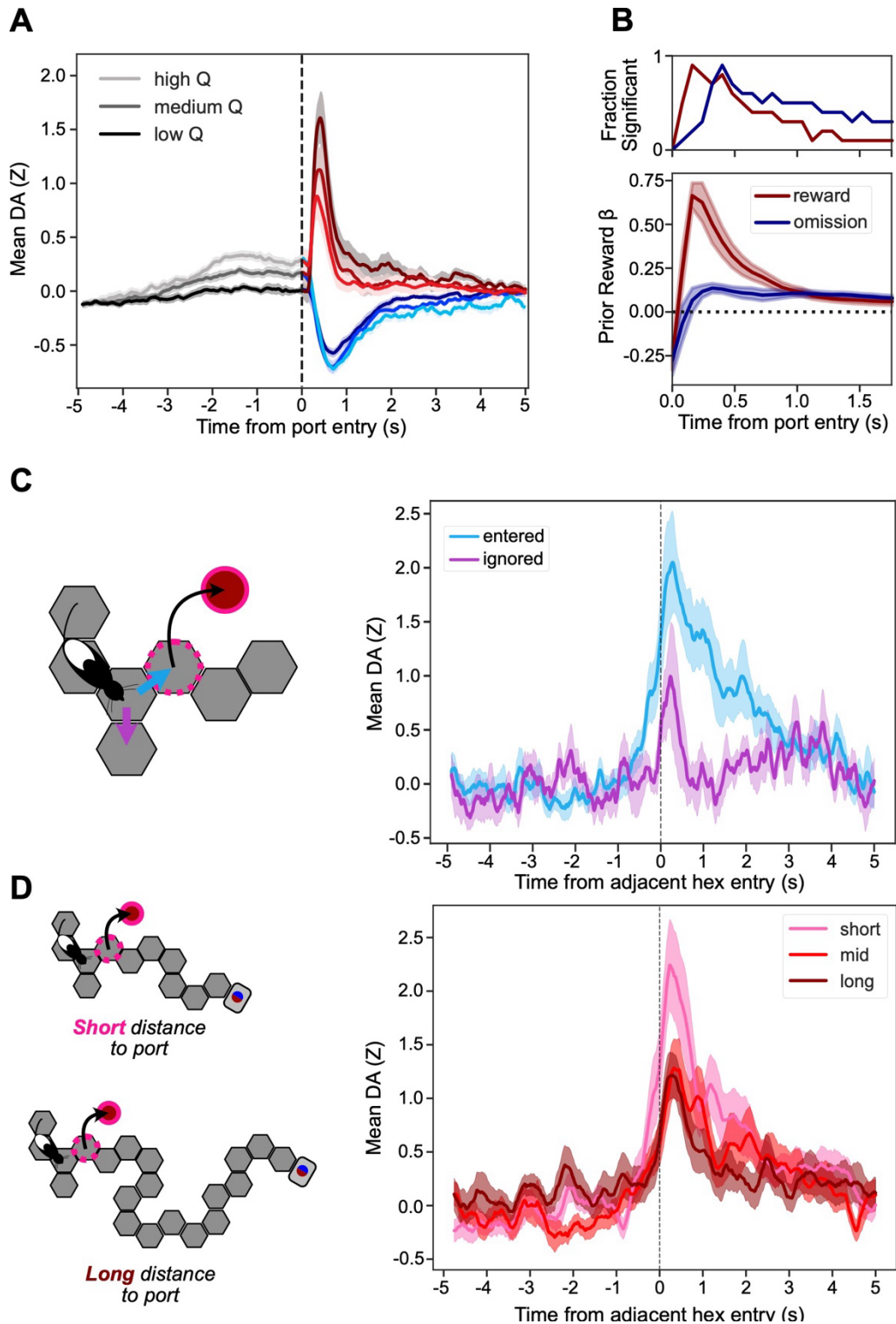
Quantification and Statistical Analyses.

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 \pm SEM, unless otherwise specified in the figure legends. Inclusion criteria for specific analyses are stated in both the 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.

A**B**

Supplemental Figure 2.1. Navigational adaptations to maze configuration changes.

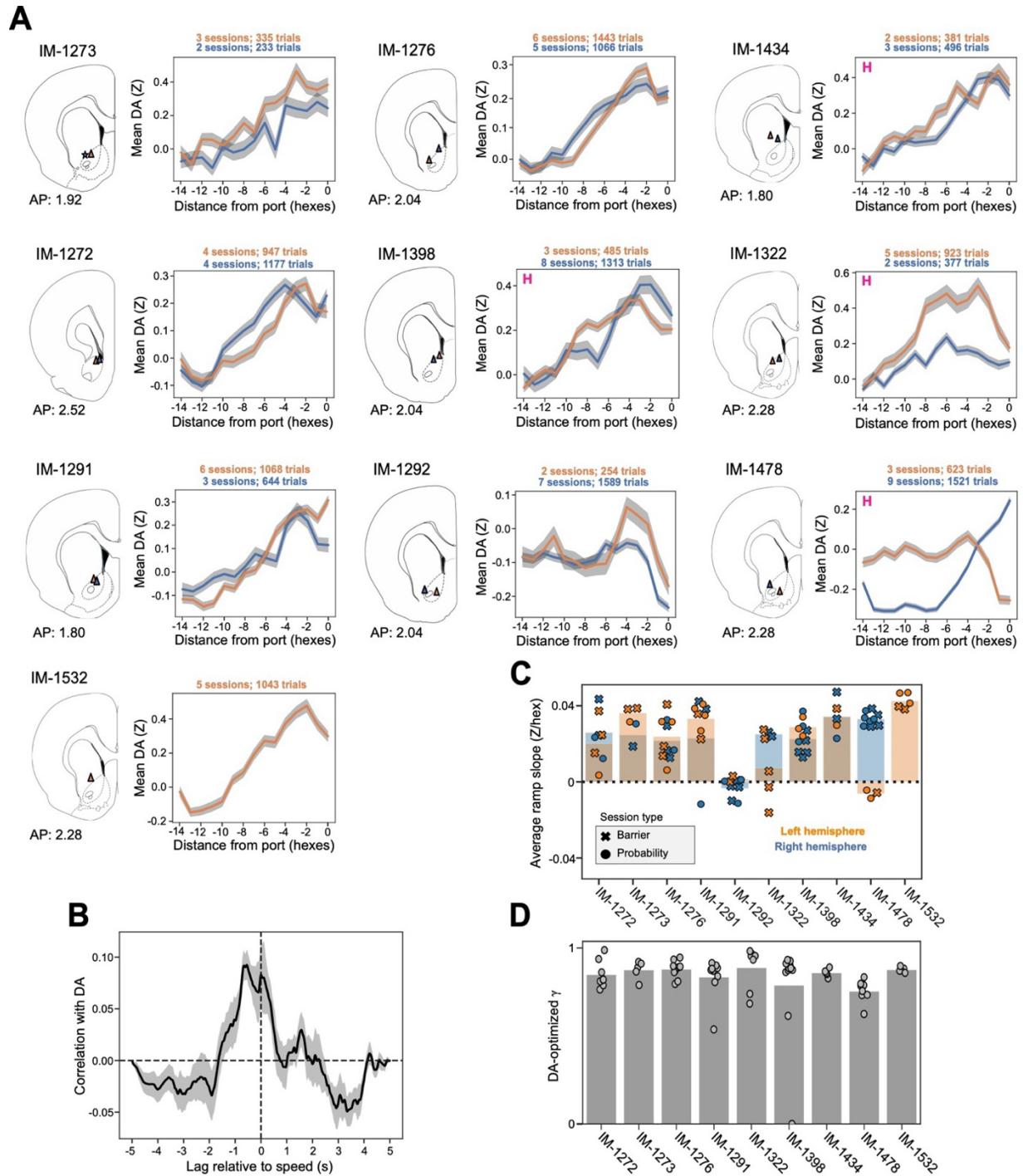
A, Position data from the first, second, and third group of ten trials following a barrier change. Green arrows highlight the progressive reduction in the distance traveled into a novel dead-end path. This barrier change is from the same block as Fig. 2.1D. B, Deviation from the optimal (shortest) path lengths over the course of the initial blocks of all sessions (*left*), blocks following a barrier change (*middle*), and blocks following a p(reward) change (*right*). Deviation is measured as the number of extra hexes traversed beyond the shortest possible path length. Dots show mean values within ten-trial bins; error bars show 95% confidence intervals. Deviation is measured as the number of extra hexes traversed beyond the shortest possible path length. Dots show mean values within ten-trial bins; error bars show 95% confidence intervals.



Supplemental Figure 2.2. Extended analysis of DA pulses.

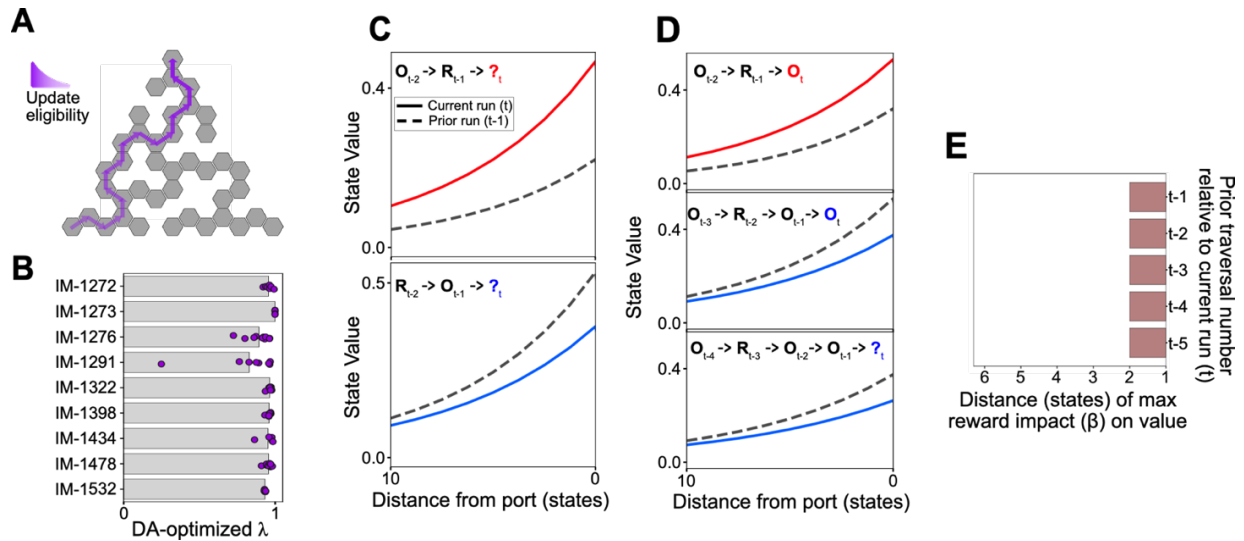
A, Port-entry aligned DA in the second half (trial number > 25) of each block ($n = 10$ rats, 82 sessions, 9,079 trials), pooled by terciles of the RL model Q value for the chosen port. B, *bottom*, regression weights for Q-value-derived RPE (see Methods) on DA following port entry (100ms bins over the first 2s). Separate regression weights are shown for RPE following reward (red) and omission (blue). (Figure caption continued on the next page).

(Figure caption continued from the previous page.) Regressions were performed independently for each rat. *Top*, fraction of rats with a significant relationship (non-zero regression coefficient, two-tailed t-test) between RPE and DA in the time bin. *C, left*, cartoon of rat choosing to enter (light blue) or ignore (violet) the newly available path. *Right*, DA aligned on entry into the hex adjacent to a newly available hex, broken down by whether the rat subsequently entered ($n = 9$ rats, 77 events) or ignored ($n = 5$ rats, 25 events) the novel path. Four rats either never chose to enter the ignored path option, or entered on fewer than three instances in total, so they were excluded from analysis. *D, left*, cartoon of a rat discovering a newly available hex with either a short or long distance to the destination port. *Right*, DA aligned on entry into the hex adjacent to a newly available hex, pooled by the tercile of distance to the destination port ($n = 10$ rats, 36 short distances, 35 mid distances, 35 long distances). All error bands indicate \pm SEM.



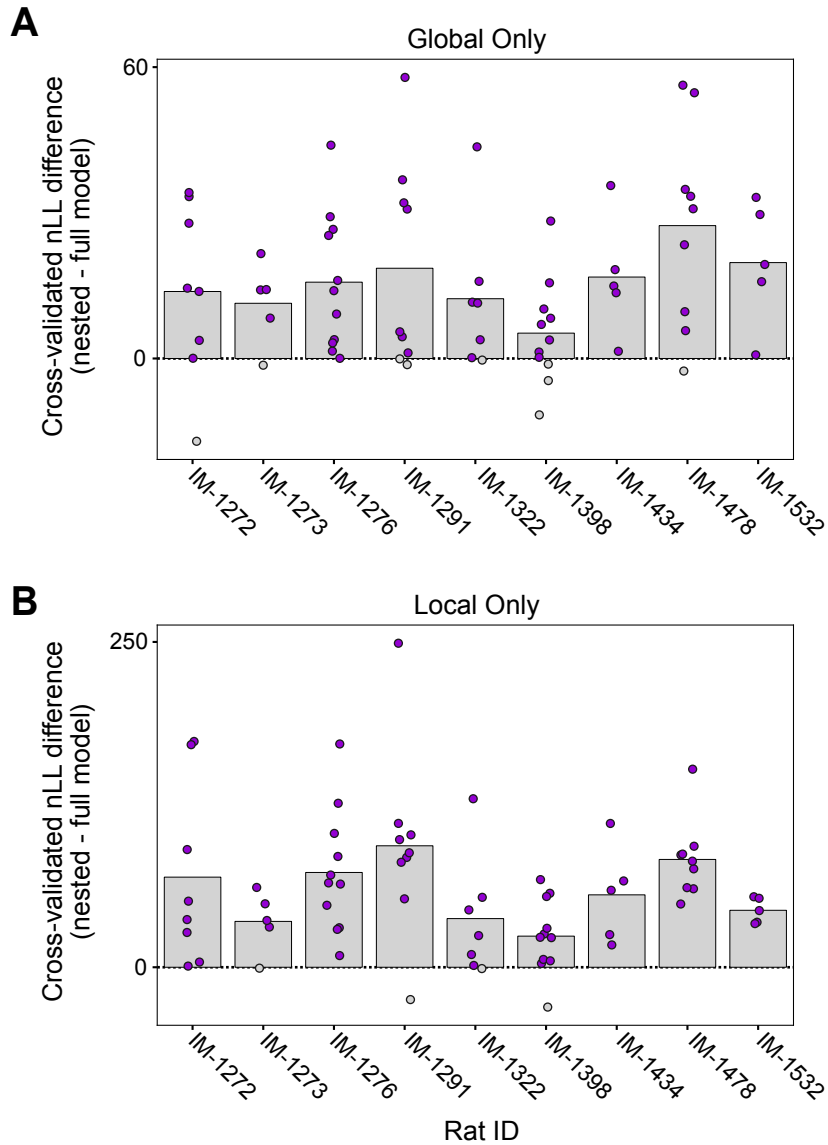
Supplemental Figure 2.3. Individual-animal recording locations and DA goal-approach ramps.

A, Histologically identified fiber locations for each animal, paired with the average ramp \pm SEM for the corresponding implanted fiber. Orange denotes right hemisphere and blue left. Blue asterisk signifies the inferred fiber location of an implant with missing histology (IM-1273 left hemisphere). Pink "H" marks rats additionally implanted with an electrophysiology probe. B, Cross-correlation between running speed and DA, aligned to running speed. C, Average ramp slope for each recording location (bar plot), with individual sessions marked as "x" for left and "o" for right hemisphere. D, Discount factors (γ) for each recording location with a significant ramp, fit to the observed DA data in the value-iteration algorithm. Bars show rat means, and dots show session values.



Supplemental Figure 2.4. Further comparison of TD algorithms to DA signals.

A, Schematic of the TD(λ) update algorithm showing the traversed hex-states' eligibility at the end of a trial. B, Estimated λ parameter values after fitting TD(λ) to each session's hex-level DA signal (bars show mean values for each rat, dots signify individual-session estimates within rats). C, Predicted DA value traces based on parameter values from the fitted TD(λ) hex-value learning algorithm. Predicted value function in response to reward and omission, over the same trial sequences as in Fig. 2.3C. "R" and "O" denote rewards and omissions, respectively, on the $t-n$ previous visits to the port. D, Predicted value function in response to a single reward in a series of omissions, over successive runs of the same path, as in Fig. 2.4B. E, Analyzing the distance from the terminal port in which prior rewards have their strongest impacts (linear regression weights) on state value, as in Fig. 2.4E/F. Predictions from 1000 simulations of a TD(λ) algorithm, based on parameter values from the fitted TD(λ) hex-value learning algorithm.



Supplemental Figure 2.5. Individual-session dual-component RL model comparison results.

A, Gain or loss of model fit to hex-level DA when the TD(0) learning component is removed from the dual-component model. Bars show mean fit comparison for individual rats, circles show model comparisons for individual sessions within rats. Positive values (purple circles) indicate superior dual-component performance (negative log likelihood, "nLL"), compared to single-component. B, Same as A, comparing when the global update component is removed from the dual-component model.

References

1. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning, second edition: An Introduction (MIT Press).
2. Schultz, W., Dayan, P., and Read Montague, P. (1997). A Neural Substrate of Prediction and Reward.
3. Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47, 129–141.
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. 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.
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.
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.

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.* 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 Nambodiri, V.M.K. (2022). Mesolimbic dopamine release conveys causal associations. *Science*, eabq6740.
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.
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. 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.
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.
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. 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.
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.
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.
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, 1–15.
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*. 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, 1–15.
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.

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. *bioRxiv*, 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.
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.
25. Namboodiri, V.M.K. (2022). How do real animals account for the passage of time during associative learning? *Behav. Neurosci.* *136*, 383–391.
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.
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.
28. Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science* *310*, 1338–1340.
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.

30. Daw, N.D., O’doherly, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*. 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.
32. Patriarchi, T., Ryan Cho, J., 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*.
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.
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.
35. Simon, D.A., and Daw, N.D. (2011). Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *J. Neurosci.* 31, 5526–5533.
36. Daw, N.D. (2009). Trial-by-trial data analysis using computational models. *Attention & Performance XXIII*.
37. Horvitz, J.C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* 96, 651–656.

38. Bromberg-Martin, E.S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. *Neuron* 68, 815–834. 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.
40. Gardner, M.P.H., Schoenbaum, G., and Gershman, S.J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences* 285, 20181645. 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.
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.
43. Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep Exploration via Bootstrapped DQN. *arXiv [cs.LG]*.
44. Walton, M.E., and Bouret, S. (2018). What Is the Relationship between Dopamine and Effort? *Trends Neurosci.*, 1–13.

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.
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.* 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.
48. Wei, W., Mohebi, A., and Berke, J.D. (2022). A Spectrum of Time Horizons for Dopamine Signals. *bioRxiv*, 2021.10.31.466705. 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, 1–8.
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.
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*, 1–17.
52. Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440. 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.
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*.
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.
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. U. S. A.* *117*, 689–697.
57. Mattar, M.G., and Daw, N.D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* *21*, 1609–1617.
58. van Hasselt, H., Madjiheurem, S., Hessel, M., Silver, D., Barreto, A., and Borsa, D. (2021). Expected Eligibility Traces. *AAAI* *35*, 9997–10005.
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. *arXiv [cs.LG]*.

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.
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.
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*. 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.
64. Wikenheiser, A.M., and Redish, D. (2015). Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* *18*. 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.
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.

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.* 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. *Journal of Neuroscience.* 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.
70. 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.
71. 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.
72. Pitis, S. (2018). Source Traces for Temporal Difference Learning. *AAAI* 32. 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.

74. Oakes, D. (1999). Direct calculation of the information matrix via the EM. *J. R. Stat. Soc. Series B Stat. Methodol.* *61*, 479–482.

Chapter 3:

Evidence that Dopamine-encoded Values are Retrieved and Updated through Mental Simulation

Introduction

The encoding properties of the dorsal CA1 region of the hippocampus (dHip) position this region as an ideal candidate for neural implementation of model-based valuation processes. The pyramidal neurons here selectively fire in spatial receptive fields (“place fields”) as animals navigate through space.¹⁻³ Ensembles of these “place cells” form a coherent representation of an animal’s actual location within an environment, sequentially firing as animals traverse through their respective place fields (“local” place coding).⁴ Upon closer examination, multiple groups have also observed encoding of locations distinct from the animal’s actual place at compressed timescales (“non-local” place coding),⁵⁻⁷ as if simulating other possible locations in the environment.

One non-local phenomenon potentially used for MB valuation reliably occurs in concert with a slow (~8Hz) local field potential (LFP) “theta” rhythm, primarily during movement through an environment.⁸ Within a single theta cycle, place cells fire in a temporally coordinated manner so that the place representation typically “sweeps” from behind the animal to locations ahead of the animal.^{6,9-11} The represented distance ahead of the animal can vary depending on current task demands,^{12,13} including alternating sweeps into possible future paths when rats approached a decision point.^{14,15} Such prospective theta sweeps would be a powerful mechanism to leverage an internal model of the environment, simulate possible future trajectories, and

evaluate them for downstream decision processes. Consistent with this hypothesis, one study found that prospective representations are associated with an MB decision-making strategy,¹⁶ and another found that MB planning in humans was hippocampus dependent.¹⁷ To date, however, no study has tested whether theta sweeps are used for prospective place value retrieval.

Theta sweeps are not the only mechanism by which the dHip might implement MB valuation. During periods of rest, non-local representations of different locations within the environment have been observed to occur in concert with a high frequency (120 to 250Hz) LFP oscillation called a sharp-wave ripple (SWR).^{5,7,18} These non-local representations include continuous trajectories,^{19,20} as well as fragmented “jumps” to distant locations within an environment.^{21,22} Consistent with an MB value-update process, SWR-associated representations include both directly experienced trajectories as well as alternatives throughout the same environment.^{23–26} Such simulations would be an ideal candidate to assign credit over space, across both previously taken and alternative trajectories to rewarding places. For example, while a rat consumes a reward, replaying distant locations may help to associate those places with downstream reward. In fact, SWRs are most likely to occur following reward receipt,²⁷ where trajectories most often start or end at the rewarding location.²⁸ For these reasons, awake replay has long been hypothesized to update value estimates.^{5,29,30} As evidence for a role in learning, selectively inhibiting SWRs impedes task learning,³¹ while optogenetically extending their duration augments task learning.³² Whether non-local representations during post-reward SWRs are used to update the values of those places is a critical gap in our understanding of how SWRs contribute to learning.

To influence decision processes, the dHip cannot act in a vacuum. We hypothesize that dHip non-local representations coordinate with downstream neural value representations. For

this, we turn to the nucleus accumbens (NAc), the ventral region of the striatum implicated in value estimation^{33–38} and decisions to pursue a goal.^{39,40} We have recently shown that dopamine (DA) signaling in the NAc scales with dynamic value estimates across space, as if forming an internal spatial map of value (see Chapter 2).⁴¹ These “place values” were updated, in part, through inference, reflecting an update mechanism that can associate rewards with both directly experienced and alternative paths.

Non-local representations could provide an ideal neural substrate to implement such inferential valuation. Consistent with cooperation between these brain regions during behavior, certain spatial navigation tasks that depend on dHip⁴² also rely on the NAc.^{43,44} Moreover, disrupting DA signaling in the NAc inhibits a rat’s ability to flexibly navigate to familiar goals from novel starting points,³⁹ a posited function of dHip non-local activity.⁴⁵ Direct projections have been identified between the dHip and the NAc,⁴⁶ and striatal units have been observed to phase-lock to dHip theta in a reward-dependent manner.^{47–49} In addition, SWRs have been associated with reactivations of putative DA neurons⁵⁰ and NAc projection neurons.^{15,51}

Thus far, however, experimental tasks and recording methods have proven difficult for testing whether neural value representations are retrieved during prospective theta sweeps or updated during SWRs. The behavioral paradigms used to study dHip non-local coding most often consist of simplified mazes with static reward contingencies and spatial configurations. This is quite useful for reducing sources of variability that make it more difficult to study non-local representations, but prohibitive for dynamic value analyses. Conversely, paradigms that have studied valuation processes often leverage spatially reduced operant tasks, where the underlying state space is ambiguous and poorly suited for non-local encoding analyses. Non-local representations, both prospective theta sweeps¹⁴ and SWRs,²⁸ are also task-dependent,

decreasing in frequency as behaviors become stereotyped. Therefore, the ideal task for this experiment would discourage sustained periods of stereotypy and encourage MB decision processes. Achieving robust decoding of non-local activity also requires the simultaneous recording of numerous pyramidal neurons,^{15,21,45} which typically requires many recording electrodes in dHip.

To overcome these obstacles, we developed a novel spatial decision-making paradigm and recording method. The maze task consists of changing paths, dead-ends, and unstable probabilistic reward contingencies. Rats must also sequentially traverse through numerous decision points in order to obtain reward, where multiple paths can lead to the same rewarding location. This task has recently allowed us to carefully study dynamic value computations in NAc DA,⁴¹ where rats were shown to exhibit MB choice behavior. To record from sufficient dHip neurons for robust non-local decoding, we employed a custom 256-channel electrophysiology probe. In the same animals, we also recorded from NAc DA, using dLight fiber photometry, to analyze the associated map of place value. We observed frequent non-local activity in the maze: both prospective theta sweeps into possible future places, and post-reward representations of places throughout the maze during SWRs. Preliminary analyses revealed that, during running through the maze, NAc DA scaled with the values of prospectively represented places. Finally, we found that post-reward non-local representations update the associated DA place values.

Results

A novel approach to simultaneously record large numbers of dHip units with NAc DA.

We implanted rats ($n=2$) with a custom 256-channel silicon electrophysiology probe, placed on a custom microdrive to target the CA1 pyramidal layer of the dHip (dCA1), along with optic fibers in the NAc for dLight fiber photometry recordings of DA (Fig. 3.1A/B). Once the silicon probe had reached dCA1 (identified using LFP signatures⁵²), we simultaneously recorded from dCA1 units and NAc DA as rats foraged for reward in the maze task. Sessions with fewer than 30 hippocampal units were excluded from the dataset, along with rats with no dLight expression ($n=1$). We were able to record from hundreds of dHip units during a single behavioral session (Fig. 3.1C), many of which displayed place-selective firing profiles in the maze (see Fig. 3.1E for examples).

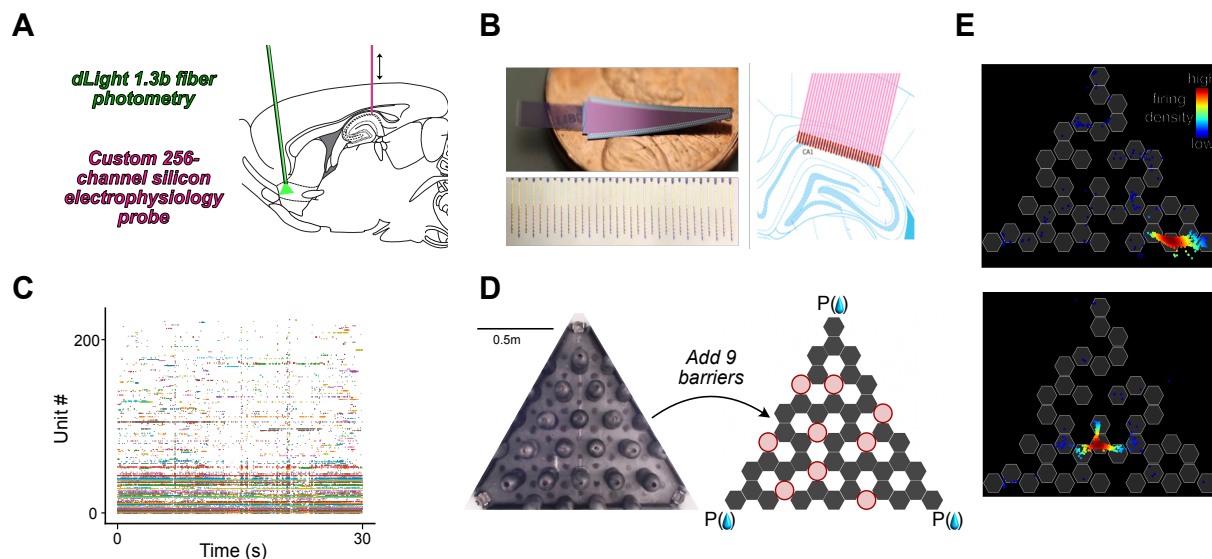


Figure 3.1. Experimental approach and place recording in a complex maze task

(A) Implant schematic. Green indicates the angled fiber photometry implant to measure dLight 1.3b. Purple shows the electrophysiology implant, placed on a movable microdrive for precise targeting along the dorsal-ventral axis. (B) Custom electrophysiology probe. Left top, magnified image of probe, against a penny for scale. Left bottom, magnified image of the 32 shanks, each with 8 recording sites. Right, cartoon emphasizing probe's design that allows for simultaneous positioning of numerous shanks in the CA1 pyramidal layer. Dark pink indicates locations of recording sites, spanning $105\mu\text{m}$. (C) Example raster plot showing spiking activity from >200 identified units in a single recording session. (D) Maze task. Left, birds-eye view of the maze, before addition of movable barriers. Right, schematic of one possible maze configuration after the addition of the nine movable barriers. Probabilistic rewards are delivered at each apex of the triangle. (E) Spatial firing profiles of two example units from a single recording session. Dots mark each location where a spike was observed, color coded by point density.

While running through the complex maze task, dHip place representations sweep ahead of the animal into possible future locations.

The maze, described previously,⁴¹ is triangular with a reward port at each corner, each with a distinct reward probability (**Fig. 3.1D**). 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 – the end location for each “trial” is the start for the next – and each reward port can be approached from multiple starting locations. After each block of 50-70 trials (traversals between ports), either the reward probabilities change or a barrier is moved to alter available paths (**Fig. 2.1**). We have previously demonstrated that rats choose ports with higher probabilities of reward and shorter distance costs.⁴¹

Because we designed this task, in part, to encourage a decision-making strategy that uses mental simulations, we first asked whether prospective theta sequences could be seen in the maze. To decode the spatial representation from dHip neural activity, we used a recently described millisecond-resolution decoder²¹ and adapted this to two-dimensional space. While rats ran through the maze (>10cm/s), we observed representations that swept ahead of the animal in coordination with the theta rhythm (**Fig. 3.2A/B**). As rats approached the numerous choice points in the maze, place representations frequently swept ahead into available prospective paths, as if simulating upcoming possibilities for evaluation. Often, these theta sweeps alternated between different paths on successive theta cycles, consistent with deliberation and a previous report.¹⁵

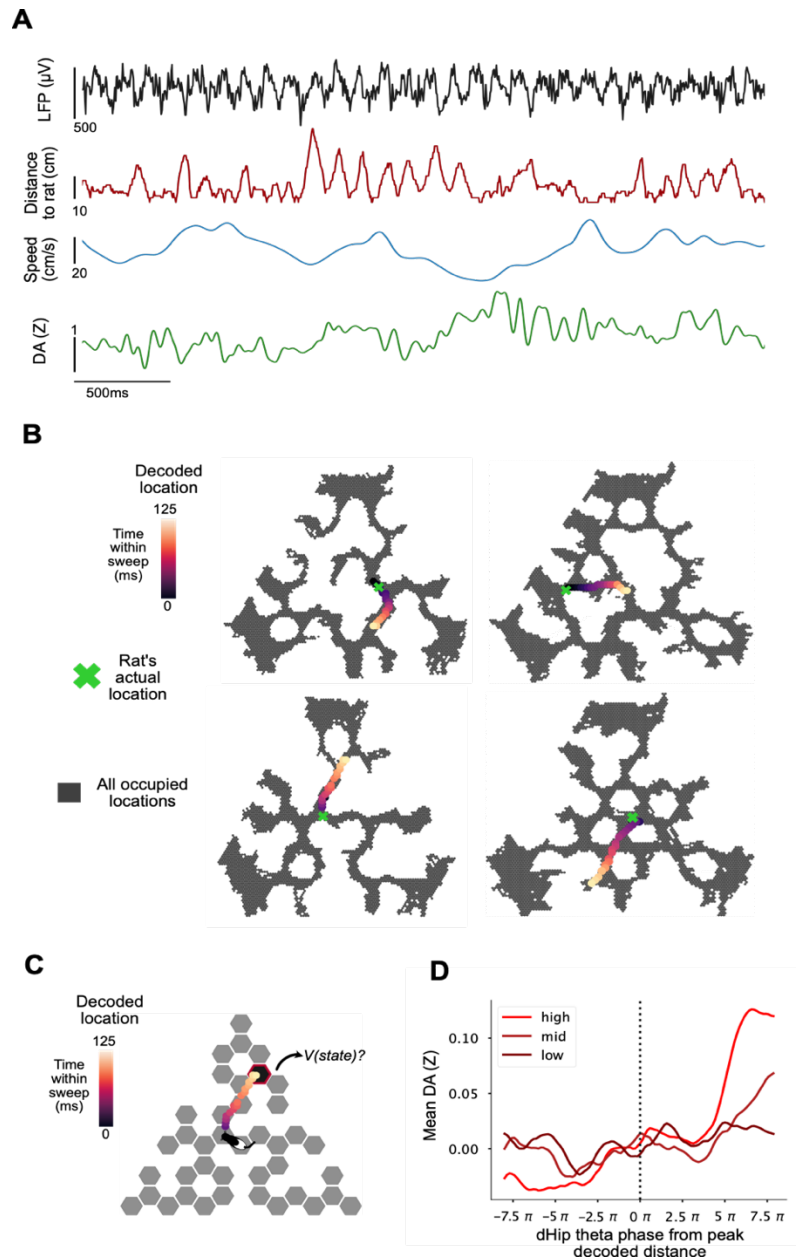


Figure 3.2. Theta-based prospective sweeps are associated with DA release proportional to the values of represented locations

(A) Example traces from one trial in the maze. From top to bottom: broadband LFP, the distance of decoded location from the rat's actual location, running speed (smoothed with a 100ms rolling mean), and z-scored dLight. Signals are all sampled at 250Hz. (B) Examples of theta-associated prospective sweeps from four different recording sessions. Green "x" marks the rat's actual location. Heat scatter plot shows the decoded location over the course of 125ms. Grey background shows all detected locations from the session. (C) Schematic of the value-retrieval analysis. The highlighted hexagon shows the location of the peak decoded distance from the rat. (D) DA aligned to the moment of peak decoded distance, pooled by the value (in terciles) of that decoded place ($n = 10$ sessions, 8399 sweeps). Traces show the average over rats, where the average over each session was first calculated for each rat. Time, here, is binned by theta phase (14 bins per cycle).

Theta-associated non-local representations modulate NAc DA to retrieve values of possible future locations.

If dHip place signals coordinate with NAc DA value signals during navigation, we would expect NAc DA fluctuations to be influenced by the spatial content encoded in dHip. We tested whether prospective representations during running influence NAc DA place values. To do this, we first identified moments when the place representation extended into possible future locations (>15cm from the rat), and then isolated the terminal locations of each sweep: the hex where the decoded representation was farthest from the animal (**Fig. 3.2C**; see Methods). We then extracted the value of each decoded terminal place using a trial-by-trial RL algorithm, which propagates up-to-date reward information from the two available ports over the current spatial map, discounted by distance (see Methods). If NAc DA represents the retrieved values of the prospective locations, we would expect DA release following the terminal place representation to scale with that place's value. Indeed, preliminary analysis reveals that when higher valued places were represented, DA release was greater than when lower valued places were represented (**Fig. 3.2D**).

Following reward receipt, dHip non-local representations update DA value representations.

While thus far this study has focused on NAc DA value dynamics during navigation, it is important to note that DA also encodes a teaching signal hypothesized to coordinate with dHip representations. Following reward, DA transiently increases in proportion to reward prediction error (RPE). These are thought to guide learning by updating values and choice behavior. It has, therefore, been hypothesized that this post-reward DA pulse might co-occur with post-reward SWRs to update values of the SWR-represented places, as if signaling the degree to which

represented locations should be updated.⁵ To test the feasibility of this hypothesis, we isolated times where SWRs were observed (**Fig. 3.3A**; see Methods) and compared the time courses of DA release versus SWR probability (**Fig. 3.3B**). Inconsistent with the co-occurrence hypothesis, we found that the DA pulse is closely time-locked to reward receipt (~200ms later) in a stereotyped manner, while SWRs do not begin until ~2s later and are less temporally precise. Nonetheless, we did reproduce prior findings that SWRs were more likely following reward compared to omission. Intriguingly, SWR probability appeared to scale with RPE on rewarded

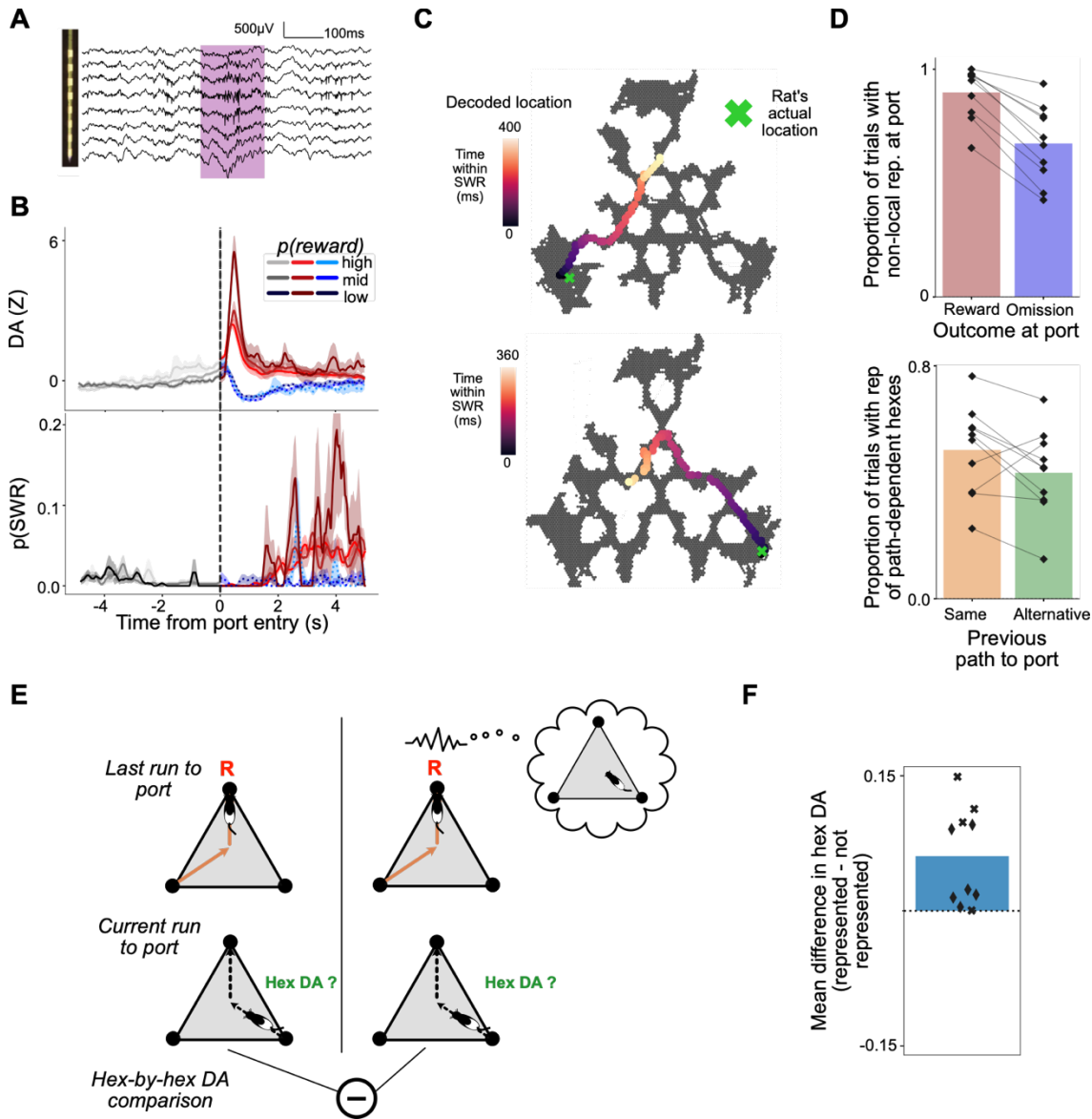


Figure 3.3. Post-reward representations follow DA reward responses and update place values

A) Wide-band LFP across the eight channels of an example recording shank. An identified SWR is highlighted in purple. (B) Reward-port-entry aligned traces. Top, mean DA pooled by reward probability at the destination port. Red indicates rewarded trials, blue indicates trials where no reward was received. Only the second halves of blocks (>25 trials) were included. Bottom, mean ripple probability, pooled by port reward probability. Color scheme is the same as above. Error bands denote s.e.m.. (C) Example decoded location during SWRs. Examples are from separate recording sessions. Green “x” marks the rat’s actual location. Heat scatter plot shows the decoded location over the course of an SWR. (D) Top, proportion of reward and reward-omission trials with identified representations of at least one non-local place. Bottom, of the trials with identified non-local representations into path-dependent hexes, proportion that entered the same path as previously taken vs alternative maze segments. Diamonds show mean values for individual sessions. Bars show mean values over sessions. (E) Illustration of the hex-level DA comparison used in analysis of post-reward decoded location. Following a reward, dHip either represents the bottom right hex or does not. DA during the following run through that hex towards the rewarded port is compared under the two conditions. (F) Average hex-by-hex difference in DA when the hex was previously represented – when that hex was not previously represented. Individual points show session averages, with different markers denoting different rats. Bar shows mean difference across sessions.

trials.

When examining the places represented following port entry, we find the non-local locations are consistent with an inferential update process. Because SWR identification involve arbitrary assumptions (see Methods), these exclude many time points from analysis when non-local places may have been represented by dHip but our algorithm failed to identify an SWR. Therefore, for the following in-depth analyses of non-local representations following port entry, we included all data that met decode-quality criteria (see Methods). Similar to the SWR analysis above, we found that representation of any non-local place in the maze was more likely following reward, compared to omission (**Fig. 3.3D**). We then asked, on the rewarded trials, whether the places represented were restricted to those along the path the rat actually took, or also extend into alternative paths to the same rewarding location. To constrain this analysis to the hexes not shared between paths, we exclusively analyzed those hexes prior to the rat's first left-right choice between reward ports ("path-dependent hexes"). Of the trials where non-local representations entered path-dependent hexes, representations were not restricted to the hexes along the previously taken path. Instead, hexes along alternative paths to the same goal location were also represented (**Fig 3.3D**). Such representations along places that both directly led to the port as well as places that *could have* led to the port would be a powerful mechanism to infer value over space.

Finally, we dove deeper to our dataset to directly test whether the non-local places represented in the dHip following reward are used to assign credit to those places, updating their values for future maze traversals. We previously showed that a single reward is able to increase DA place values across an entire path on the subsequent run through to that port.⁴¹ Can this spatial value propagation be explained by dHip representations of those distant places following

reward? To explore this possibility, we first identified all trials where the rat received reward at a destination port. We then focused our analysis on the rat's subsequent traversal to that same port. If non-local representations act to update place values, DA in a specific hex should be higher when that hex was previously represented, compared to the runs where it was not. We performed this exact analysis (**Fig. 3.3E**, see methods), and we found that for each traversed hex, on average, DA was higher if that hex was previously represented versus reward alone (**Fig. 3.3F**). Though preliminary, the observed effect is consistent across all 10 sessions recorded from two separate rats. This provides strong evidence supporting the hypothesis that SWRs influence behavior by updating the values of distant states.

Discussion

While preliminary, this study is the first to present direct evidence for a neural circuit that leverages mental simulation to infer upcoming reward. During running, dHip representations of possible future locations resulted in downstream increases in NAc DA, which scaled with the values of the represented places. The increase in DA occurred approximately 200ms following the prospective place representation. This delay in NAc DA response is similar to what we⁵³ have previously observed when rats were given a sensory reward-predictive cue. Such a delayed value representation is likely not used for directional action selection. Directional action values may instead be represented in the spiking of orbitofrontal or prefrontal cortical neurons,⁵⁴⁻⁵⁷ or striatal spiny projection neurons.^{38,58,59} Combined recordings of dHip with these candidate regions are the subject of future and ongoing investigations. NAc DA is likely performing the same function as observed in other paradigms.^{34,39} signaling whether the current task is worth engaging in, compared to, e.g., resting and grooming. Consistent with a signal for controlling motivational vigor, we previously found that NAc DA levels predicted running speed in this task.⁴¹ Under this model, successive theta sweeps serve to sample upcoming locations, where NAc DA collects evidence to evaluate whether the rat should continue running through the maze. To better test this hypothesis, one avenue for future study in the maze task is whether NAc DA *decreases* when locations of null value, such as a dead-end, are simulated. If NAc DA is indeed used to assess whether a goal is worth pursuing, such a dip in the signal should be associated with the decision to abandon a dead-end path or avoid it altogether. It also bears mentioning that while we interpreted this NAc DA response as a value signal, DA may actually compute RPE here. Simulation-evoked transients may only occur when the place represented is better or worse than expected. Further modeling work is necessary to tease apart the two interpretations, but it's worth

noting that our analysis assessed DA as a function of the difference between simulated and local place value. If simulated values act as the “observed” value to be compared with the local “expected” value, then the DA transient may indeed reflect an RPE signal.

We also found evidence that mental simulations are used to implement credit assignment over space, bridging gaps between observed rewards and distant locations. Following reward receipt, dHip representations of distant places resulted in higher NAc DA on the subsequent run through those places, compared to reward alone. This is the first direct evidence, to our knowledge, for a neural circuit that uses simulation to update downstream representations of value. This mechanism would be a powerful tool for agents to adaptively update value estimates in naturalistic tasks, where routes to reward are both unstable and numerous. It also may explain the inferential valuation we observed previously.⁴¹ We plan to directly test this hypothesis by assessing whether the value-learning RL model used in that paper can better predict NAc DA if global updates are restricted to the set of places represented by dHip following reward.

Still, it remains unclear how simulating a distant place can update its value. In contrast to the model hypothesized by Foster and Wilson,⁵ we did not observe concurrent SWRs with DA reward responses. Instead, dHip simulations may be used, along with reactivated downstream representations^{60–65} to implement an offline-learning algorithm similar to Dyna Q learning.^{66,67} This RL algorithm simulates states within the current environment and learns as if traversing those states in reality. Perhaps, by reactivating distant place representations and those associated with the rewarding state in close temporal succession, the distant places can develop an association with downstream rewarding states. Understanding whether the neural mechanism for these changes operates through plasticity⁶⁸ or changes in the population dynamical state⁶⁹ will require further investigation.

It is perhaps more than coincidence that both neural signals under investigation operate under two distinct encoding modes: one associated with running and reward seeking, the other consumption and rest. As described previously,^{34,35,41,70,71} NAc DA increases in a ramp as rats approach rewarding locations, and it scales with value. However, this ramp becomes disengaged when behavior becomes stereotyped.^{71,72} Non-local theta sweeps have similarly been shown to decrease in prevalence as rats' experience with a task increases and stereotyped behaviors are more likely.¹⁴ Future analyses of the current dataset should test whether prospective theta sweeps are predictive of NAc DA ramping, indicating a coordinated information-processing mode between the two regions. This would not be surprising, as coordination between dHip and VTA DA neurons was observed in a simpler maze task, but only during the cognitively demanding portion of the track.⁷³ Following reward, DA switches to transiently report a learning signal, RPE,^{41,74} while dCA1 decreases theta power and sporadically engages SWRs.^{52,75} It is, therefore, possible that the brain enters a learning mode during rest and consumption, distinct from the online planning-and-evaluation mode during running and reward seeking.

Related to the switch between different encoding modes, the brain must also perform meta-decision-making to choose which places should be represented at any given moment. Why do dHip representations sweep into both available paths at most choice points, but not all? It is also a mystery why some theta sequences are longer than others, and what meta decision process determines when to stop sweeping ahead. Relatedly, which places within sometimes large environments should be represented during SWRs? One prominent hypothesis²⁹ posits that places should be represented based on their need – how frequently the animal is expected to visit the place – and gain – the additional reward expected if the place's value is updated and subsequent behavior changed. This model has successfully replicated experimental observations

associated with SWRs, but our dataset could help answer whether the same model can explain the places represented during theta sequences. It's worth noting that SWRs are also observed when animals pause as they navigate a maze, and these are sometimes, but not always,²² associated with prospective representations related to behavior.^{7,26,45} The prominence and function of these events were not investigated in this study, but they are the subjects of further analyses on the current dataset.

Methods

Animals. All animal procedures were approved by University of California San Francisco Institutional Committees on Use and Care of Animals. Male (300–650g) wild-type Long-Evans rats (4-8 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.

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 15 μ L 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. Rats’ implant caps were labeled and tracked using Deeplabcut.⁷⁶ Custom code was used to segment the maze into hexes and classify hex occupancy.

Neural Recordings. 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 – 1 μ L of AAVDJ-CAG-dLight1.3b (Vigene) at a titer of 2×10^{12} – was delivered

using a stereotaxic injection pump (Nanoject III). Virus was injected 200 μ m ventral to the target coordinates, as described in.³⁵ During the same surgery, 200 μ m optical cannulae were subsequently implanted and cemented in place. We then performed a unilateral craniotomy above the dorsal hippocampus, centered at 4.3mm posterior and 2.2mm lateral, relative to bregma. A 2.5mm wide (coronal plane) durectomy, centered at 2.2mm lateral, was then performed in the region between 3.6-4.5mm posterior that best avoided blood vessels. The probe was subsequently inserted into the exposed brain, to a depth of 1.5mm-2.00mm below the brain. A screw was implanted above cerebellum to serve as a global reference. 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. In a subset of rats, the photometry implant was performed in a separate surgery at least two weeks prior to the electrophysiology implant, to ensure dLight expression before performing the subsequent implant.

To target dCA1 pyramidal cells, electrophysiology measurements were performed daily following surgery, using well-described LFP signatures to localize the probe.⁵² These recordings were performed outside of the maze, in a separate recording chamber. Following each recording, the custom microdrive was used to lower probes a distance between 40 μ m and 80 μ m. The shank geometry of the probe allowed for clear observation of probe movement in relation to spiking units. To account for possible delays in probe movement following driving – e.g., from coagulated tissue at the brain-probe interface – the probe was only lowered once every four hours at most. Driving was terminated once the slow frequency “wave” component of the SWR was observed to flatten and flip polarities on the ventral recording sites, a clear indication that

recording sites are in CA1. The probe was lowered to increase the number of shanks displaying this LFP signature to the greatest extent possible, to maximize the number of units recorded from the pyramidal layer. After rats had recovered from surgery (~1 week), rats were retrained in the maze task (alternating between barrier and probability variants) until sufficient recording sites reached their target location.

Photometry data acquisition methods have been described previously.³⁵ Baseline correction was performed using the adaptive iteratively reweighted Penalized Least Squares (airPLS) algorithm.⁷⁷ 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. The optical fiber was passed through a through-hole electrical commutator, which was used for the electrophysiology recordings. Recording locations were histologically verified using immunohistochemistry.³⁵ 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. To capture higher frequency fluctuations in the signal that might coordinate with dHip, further smoothing was not performed. 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.

Electrophysiology signals were acquired at 30KHz, using the RHD recording controller from Intan Technologies. Data was recorded with the open-source Open Ephys data acquisition software. Spike sorting was performed using the fully automated Mountainsort software.⁷⁸ To remove sources of noise beyond the algorithm's automated denoising steps, we excluded all recording sites whose impedance was below 200KOhms or above 3.5MOhms, indicative of shorted or broken sites, respectively. All identified clusters were included in the spatial encoding model, described below.

Theta-based analyses. For LFP analysis of the theta rhythm, a channel dorsal to the CA1 pyramidal layer and ventral to corpus collosum was selected on each recording session. LFP was filtered between 5 and 11Hz. Theta phase was computed through Hilbert transform. For theta-event-aligned plots of DA signals, photometry samples were first binned by theta phase (14 bins). Moments of peak decoded distance from the rat were defined as local maxima that exceeded 15cm from the rat's current location (head position), occurred while the rat was running ($>10\text{cm/s}$), and had a highest posterior spatial density (HPD;²¹a decode confidence measure quantifying the area containing 95% of the posterior distribution) below 28cm^2 . For each of these detected moments, a phase-binned DA trace was extracted, starting four cycles prior and extending four cycles following the moment of peak decoded distance. DA traces were

then sorted by the value of the decoded place (hex value; described below) – the value of the rat's current location. Traces were then pooled by tercile of this relative value estimate.

SWR detection. SWR times were detected using a method adapted from those previously described.^{79,80} On each shank with identified channels in the CA1 pyramidal layer, one channel with the most spikes in the pyramidal layer was selected for analysis. Broadband LFP was filtered between 150-250Hz, squared, and then summed across all identified channels. For SWR analysis, the square root of the Gaussian-smoothed summed signal was used. SWR times were identified as those when this signal exceeded 2s.d. of the mean for at least 15ms, where start and end times were defined by the prior and subsequent moments that the trace returned to the mean. Only moments where rats were immobile (<4cm/s) were included in analysis.

Post-reward place-value update analysis. We first identified which places were represented in the dHip on each trial, following port entry. The post-port-entry epoch was defined as the time between the initial entry into a port, and the first moment when the rat was detected entering a non-port-adjacent hex. We also restricted the analyzed times to periods where running speed was below a threshold (4cm/s). Non-local places were defined as those at least 30cm from the rat's current location, a distance that captures the extent of a rat's body and tail. To exclude noisy decoded place estimates with low confidence, decoded places were included for analysis if multi-unit firing rate exceeded a threshold (5Hz) and the HPD was below a threshold (28cm²). Accepted decoded places were then converted to hex ID. We did not exclude periods outside of identified SWR epochs, because non-local representations can extend beyond these experimenter-defined boundaries.

We also 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, value representations 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 visits to the destination port, we only included trials where that same port was visited and rewarded exactly two trials prior (when the port was visited at the last opportunity).

For each hex traversed over a rat's path to a specific goal port, the associated NAc DA was stored according to one of two conditions: when that hex had been previously represented, or when no representation of that hex occurred during the prior goal-port visit. For each hex, the average DA was computed under each of the two conditions in a given session. In each traversed hex, the mean DA in that hex for the no representation condition was subtracted from the mean DA for the representation condition. For each session, the mean of these subtracted values over all traversed hexes was then computed. This hex-by-hex comparison allowed us to control for differences in spatial location and distance from the reward ports, which we know have strong effects on the DA signal.⁴¹

Reinforcement Learning Models. To estimate trial-by-trial hex values based on rats' experienced rewards, while respecting maze geometry, we used a value-iteration algorithm similar to that described previously.⁴¹ We first specified ground truth hex-state transition matrices for each unique maze configuration. A 49-hex state space was used, in order to avoid a directional requirement and include a greater number of decoded places. On each trial, the values of the two available ports were set to their respective $Q(\text{port})$ values, and only the last-visited

port was set to a value of zero. The same three-port Q-learning algorithm described before⁴¹ was used to estimate dynamically evolving Q(port) values over trials. 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. This was repeated for each trial. The update rule took the following form:

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

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 function⁸¹).

Spatial Decoding and Trajectory Categorization. We used a state space model to decode the "mental" spatial position of the animal and whether the movement was consistent with a spatially continuous or fragmented trajectory from clustered spiking data.^{21,22,82,83} This model has two latent variables:

1. x_k , a continuous latent variable corresponding to the 2D position - (x_k, y_k) - represented by the population hippocampal spiking activity at time t_k .
2. I_k , a discrete latent variable corresponding to the type of movement model: spatially continuous or fragmented.

Our model estimates $p(x_{1:T}, I_{1:T} | \Delta N_{1:T}^{1:C})$, the posterior probability of position and movement model given the spikes $\Delta N_{1:T}^{1:C}$ from all cells C from time 1 to T.

We estimate this as we described previously²¹ by applying a recursive causal filter according to the following equation:

$$p(x_k, I_k | \Delta N_{1:k}^{1:C}) \propto p(\Delta N_k^{1:C} | x_k, I_k) * \sum_{I_{k-1}} \int p(x_k | x_{k-1}, I_k, I_{k-1}) Pr(I_k | I_{k-1}) p(x_{k-1}, I_{k-1} | \Delta N_{1:k-1}^{1:C}) dx_{k-1}$$

and then applying an acausal smoother starting from the last estimate of the acausal filter $p(x_T, I_T | \Delta N_{1:T}^{1:C})$ and iterating back to time 1:

$$p(x_{k+1}, I_{k+1} | \Delta N_{1:k+1}^{1:C}) = p(x_k, I_k | \Delta N_{1:k}^{1:C}) * \sum_{I_{k+1}} \int \frac{p(x_{k+1} | x_k, I_{k+1}, I_k) Pr(I_{k+1} | I_k)}{p(x_{k+1}, I_{k+1} | \Delta N_{1:k}^{1:C})} p(x_{k+1}, I_{k+1} | \Delta N_{1:k+1}^{1:C}) dx_{k+1}$$

where:

$$p(x_{k+1}, I_{k+1} | \Delta N_{1:k}^{1:C}) = \sum_{I_k} \int p(x_{k+1} | x_k, I_{k+1}, I_k) Pr(I_{k+1} | I_k) p(x_k, I_k | \Delta N_{1:k}^{1:C}) dx_k$$

We approximated the integrals over position using Riemann sums by discretizing the position space (Δ_x, Δ_y) into 2 cm bins in the x- and y-direction. We used a timestep Δ_k of 4 ms.

We defined the initial conditions of the model $p(x_0, I_0) = Pr(I_0)p(x_0 | I_0)$ as equiprobable over discrete movement models $Pr(I_0) = 0.5$ and uniform over the position state space $p(x_0 | I_0) =$

$\mathcal{U}(\min x, \max x)$ of positions the animal had visited during the epoch. This was done to avoid biasing the model towards a particular movement model or position.

We defined the discrete movement model transition probabilities $Pr(I_k | I_{k-1})$ as a 2x2 matrix with the probability of staying in the same movement model as 0.968 and the probability of switching movement models as 0.032. This was chosen to match the expected duration of half theta cycles.

We defined the movement models $p(x_k | x_{k-1}, I_k, I_{k-1})$ for each discrete transition as follows: For staying in the spatially continuous state, we defined the movement model as a 2D Gaussian random walk with mean x_{k-1} and covariance matrix Σ_{k-1} , where Σ_{k-1} has a variance of 12.0 cm and a correlation of 0.0. For staying in the spatially fragmented state, we defined the movement model as Uniform over the position state space $p(x_k | x_{k-1}, I_k, I_{k-1}) = \mathcal{U}(\min x, \max x)$ of positions the animal had visited during the epoch. For transitions between the discrete states, this was also set to spatially uniform.

We defined the population likelihood of spiking $\Delta N_k^{1:C}$ given the position and movement model $p(\Delta N_k^{1:C} | x_k, I_k)$ as a Poisson distribution:

$$\prod_{i=1}^C [\lambda_i(t_k | x_k, I_k) \Delta_k]^N \exp[-\lambda_i(t_k | x_k, I_k) \Delta_k]$$

where $\lambda_i(t_k | x_k, I_k)$ is the "place field" of the cell estimated from the encoding model. We used a kernel density estimate with kernel bandwidth 6 cm to estimate the place field during the encoding period - which was the entire epoch. See Denovellis et al. 2021 for details.²¹

Most Probable Decoded Position. The most probable decoded position is the position with the highest posterior probability density at each time point k . We marginalized over the movement model I_k , and then took the maximum of the posterior probability density over the position state space to obtain the most probable decoded position.

Ahead-Behind Distance. The ahead-behind distance is defined as the distance between the most probable decoded position and the actual position of the animal at each time point k where positive indicates the most likely decoded position is ahead of the animal and negative indicates that the most likely decoded position is behind the animal. To estimate this, we first marginalized over the movement model I_k , and then computed the shortest path distance (via Dijkstra's algorithm) between the grid bin containing the animal's position and the grid bin containing the most probable decoded position. We then computed the cosine similarity between the animal's head direction and the direction of the most likely decoded position. The direction of the most likely decoded position was defined by the direction of the first edge in the shortest path starting at the grid bin containing the animal's position. We then multiplied the shortest path distance by the cosine similarity to obtain the ahead-behind distance.

Highest Posterior Density (HPD) Spatial Coverage. The highest posterior density is a measure of the spread of the posterior probability density and reflects the uncertainty of the model about

the mental position x_k . It is defined as the smallest region of the posterior probability density that contains a given percentage of the probability mass (see Denovellis et al. 2021²¹ for details). We used the 95% HPD to define the spatial coverage of the posterior probability density. We marginalized over the movement model I_k to obtain the HPD spatial coverage for each time point.

References

1. O'keefe, J. (1976). Place Units in the Hippocampus of the Freely Moving Rat.
2. O'keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.*, 171–175.
3. Liberti, W.A., Schmid, T.A., Forli, A., Snyder, M., and Yartsev, M.M. (2022). A stable hippocampal code in freely flying bats. *Nature*. 10.1038/s41586-022-04560-0.
4. Wilson, M.A., and McNaughton, B.L. (1993). Dynamics of the Hippocampal Ensemble Code for Space.
5. Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* 440. 10.1038/nature04587.
6. Foster, D.J., and Wilson, M.A. (2007). Hippocampal Theta Sequences. *Hippocampus* 17, 1–3.
7. Diba, K., and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* 10. 10.1038/nn1961.
8. Buzsaki, G. *Theta Oscillations in the Hippocampus*.
9. Skaggs, W.E., McNaughton, B.L., Wilson, M.A., and Barnes, C.A. (1996). Theta Phase Precession in Hippocampal Neuronal Populations and the Compression of Temporal Sequences. *Hippocampus* 6, 149–172.

10. Dragoi, G., and Buzsáki, G. (2006). Temporal Encoding of Place Sequences by Hippocampal Cell Assemblies. *Neuron* 50, 145–157.
11. Schmidt, R., Diba, K., Leibold, C., Schmitz, D., Buzsáki, G., and Kempter, R. (2009). Single-Trial Phase Precession in the Hippocampus. 10.1523/JNEUROSCI.2270-09.2009.
12. Wikenheiser, A.M., and Redish, D. (2015). Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* 18. 10.1038/nn.3909.
13. Gupta, A.S., Van Der Meer, M.A.A., Touretzky, D.S., and Redish, D. (2012). Segmentation of spatial experience by hippocampal theta sequences. *Nat. Neurosci.* 15. 10.1038/nn.3138.
14. Johnson, A., and Redish, A.D. (2007). Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *J. Neurosci.* 10.1523/JNEUROSCI.3761-07.2007.
15. 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.
16. Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., and Daw, N.D. (2015). Model-based choices involve prospective neural activity. *Nat. Neurosci.* 18, 767–772.
17. Vikbladh, O.M., Meager, M.R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., and Daw, N.D. (2019). Hippocampal Contributions to Model-Based Planning and Spatial Memory. *Neuron*. 10.1016/j.neuron.2019.02.014.

18. Joo, H.R., and Frank, L.M. The hippocampal sharp wave-ripple in memory retrieval for immediate use and consolidation Nature reviews | NeurosciNce. Nat. Rev. Neurosci. 10.1038/s41583-018-0077-1.
19. Karlsson, M.P., and Frank, L.M. (2009). Awake replay of remote experiences in the hippocampus. Nature Publishing Group *12*. 10.1038/nn.2344.
20. Davidson, T.J., Kloosterman, F., and Wilson, M.A. (2009). Hippocampal replay of extended experience. Neuron *63*, 497–507.
21. Denovellis, E.L., Gillespie, A.K., Coulter, M.E., Sosa, M., Chung, J.E., Eden, U.T., and Frank, L.M. (2021). Hippocampal replay of experience at real-world speeds. Elife *10*. 10.7554/eLife.64505.
22. Gillespie, A.K., Astudillo Maya, D.A., Denovellis, E.L., Liu, D.F., Kastner, D.B., Coulter, M.E., Roumis, D.K., Eden, U.T., and Frank, L.M. (2021). Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. Neuron *109*, 3149-3163.e6.
23. 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.
24. 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. U. S. A. *117*, 689–697.

25. Gupta, A.S., van der Meer, M.A.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal replay is not a simple function of experience. *Neuron* *65*, 695–705.
26. Freyja Ólafsdóttir, H., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife* *4*, 1–17.
27. Singer, A.C., and Frank, L.M. (2009). Rewarded outcomes enhance reactivation of experience in the hippocampus. *Neuron* *64*, 910–921.
28. 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*.
10.1016/j.neuron.2016.07.047.
29. Mattar, M.G., and Daw, N.D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* *21*, 1609–1617.
30. Foster, D.J. (2017). Replay Comes of Age. *Annu. Rev. Neurosci.* *40*, 581–602.
31. Jadhav, S.P., Kemere, C., German, W.P., and Frank, L.M. (2012). Awake Hippocampal Sharp-Wave Ripples Support Spatial Memory. *Science* *336*. 10.1126/science.1221443.
32. Fernández-Ruiz, A., Oliva, A., Fermino De Oliveira, E., Rocha-Almeida, F., Tingley, D., and Buzsáki, G. (2019). Long-duration hippocampal sharp wave ripples improve memory. *Science*, 1082–1086.

33. Khamassi, M., Mulder, A.B., Tabuchi, E., Douchamps, V., and Wiener, S.I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. *Eur. J. Neurosci.* 10.1111/j.1460-9568.2008.06480.x.
34. 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.
35. 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.
36. Sugam, J.A., Day, J.J., Wightman, R.M., and Carelli, R.M. (2012). Phasic nucleus accumbens dopamine encodes risk-based decision-making behavior. *Biol. Psychiatry* 71, 199–205.
37. Strait, C.E., Slezzer, B.J., and Hayden, B.Y. (2015). Signatures of value comparison in ventral striatum neurons. *PLoS Biol.* 10.1371/journal.pbio.1002173.
38. Faust, T.W., Mohebi, A., and Berke, J.D. (2023). Reward expectation selectively boosts the firing of accumbens D1+ neurons during motivated approach. *bioRxiv*, 2023.09.02.556060. 10.1101/2023.09.02.556060.
39. 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. *Journal of Neuroscience.* 10.1523/JNEUROSCI.3958-10.2010.

40. Floresco, S.B. (2015). The Nucleus Accumbens: An Interface Between Cognition, Emotion, and Action. *Annu. Rev. Psychol.* 10.1146/annurev-psych-010213-115159.
41. Krausz, T.A., Comrie, A.E., Kahn, A.E., Frank, L.M., Daw, N.D., and Berke, J.D. (2023). Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron.* 10.1016/j.neuron.2023.07.017.
42. Moser, E., Moser, M.-B., and Andersen, P. (1993). Spatial Learning Impairment Parallels the Magnitude of Dorsal Hippocampal Lesions, but Is Hardly Present following Ventral Lesions.
43. Seamans, J.K., and Phillips, A.G. (1994). Selective memory impairments produced by transient lidocaine-induced lesions of the nucleus accumbens in rats. *Behav. Neurosci.* 108, 456–468.
44. Ploeger, G.E., Spruijt, B.M., and Cools, A.R. (1994). Spatial localization in the Morris water maze in rats: Acquisition is affected by intra-accumbens injections of the dopaminergic antagonist haloperidol. *Behavioral Neuroscience* 108, 927–934. 10.1037/0735-7044.108.5.927.
45. Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79.
46. Trouche, S., Koren, V., Doig, N.M., Ellender, T.J., El-Gaby, M., Lopes-dos-Santos, V., Reeve, H.M., Perestenko, P.V., Garas, F.N., Magill, P.J., et al. (2019). A Hippocampus-Accumbens Tripartite Neuronal Motif Guides Appetitive Memory in Space. *Cell.* 10.1016/j.cell.2018.12.037.

47. van der Meer, M.A.A., and Redish, A.D. (2011). Theta Phase Precession in Rat Ventral Striatum Links Place and Reward Information. *Journal of Neuroscience*. 10.1523/JNEUROSCI.4869-10.2011.
48. Berke, J.D., Okatan, M., Skurski, J., and Eichenbaum, H.B. (2004). Oscillatory entrainment of striatal neurons in freely moving rats. *Neuron*. 10.1016/j.neuron.2004.08.035.
49. Lansink, C.S., Meijer, G.T., Lankelma, J.V., Vinck, M.A., Jackson, J.C., and Pennartz, C.M.A. (2016). Reward Expectancy Strengthens CA1 Theta and Beta Band Synchronization and Hippocampal-Ventral Striatal Coupling. *Journal of Neuroscience*. 10.1523/JNEUROSCI.0682-16.2016.
50. Gomperts, S.N., Kloosterman, F., and Wilson, M.A. (2015). VTA neurons coordinate with the hippocampal reactivation of spatial experience. *Elife* 4. 10.7554/eLife.05360.
51. Lansink, C.S., Goltstein, P.M., Lankelma, J.V., Joosten, R.N.J.M.A., McNaughton, B.L., and Pennartz, C.M.A. (2008). Preferential Reactivation of Motivationally Relevant Information in the Ventral Striatum. *Journal of Neuroscience*. 10.1523/JNEUROSCI.1054-08.2008.
52. Buzsáki, G. (2015). Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* 25, 1073–1188.
53. Wei, W., Mohebi, A., and Berke, J.D. (2022). A Spectrum of Time Horizons for Dopamine Signals. *bioRxiv*, 2021.10.31.466705. 10.1101/2021.10.31.466705.

54. Bouret, S., and Richmond, B.J. (2010). Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. *J. Neurosci.* *30*, 8591–8601.
55. Sul, J.H., Kim, H., Huh, N., Lee, D., and Jung, M.W. (2019). Distinct Roles of Rodent Orbitofrontal and Medial Prefrontal Cortex in Decision Making. *Neuron*. 10.1016/j.neuron.2010.03.033.
56. Rudebeck, P.H., Saunders, R.C., Lundgren, D.A., and Murray, E.A. (2017). Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes. *Neuron*. 10.1016/j.neuron.2017.07.042.
57. Tsutsui, K.I., Grabenhorst, F., Kobayashi, S., and Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. *Nat. Commun.* 10.1038/ncomms12554.
58. van der Meer, M.A.A., and Redish, D.A. (2009). Covert expectation-of-reward in rat ventral striatum at decision points. *Front. Integr. Neurosci.* 10.3389/neuro.07.001.2009.
59. Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science* *310*, 1338–1340.
60. Shin, J.D., Tang, W., Jadhav, S.P., Shin, J.D., Tang, W., and Jadhav, S.P. (2019). Dynamics of Awake Hippocampal-Prefrontal Replay for Spatial Learning and Memory-Guided Decision Making. *Neuron*, 1–16.

61. Tang, W., and Jadhav, S.P. (2019). Sharp-wave ripples as a signature of hippocampal-prefrontal reactivation for memory during sleep and waking states. *Neurobiol. Learn. Mem.* *160*, 11–20.
62. Pennartz, C.M.A. (2004). The Ventral Striatum in Off-Line Processing: Ensemble Reactivation during Sleep and Modulation by Hippocampal Ripples. *Journal of Neuroscience*. 10.1523/JNEUROSCI.0575-04.2004.
63. Sosa, M., Joo, H.R., and Frank, L.M. (2020). Dorsal and Ventral Hippocampal Sharp-Wave Ripples Activate Distinct Nucleus Accumbens Networks. *Neuron* *105*, 725-741.e8.
64. Lansink, C.S., Goltstein, P.M., Lankelma, J.V., Mcnaughton, B.L., and Pennartz, C.M.A. (2009). Hippocampus Leads Ventral Striatum in Replay of Place-Reward Information. *PLoS Biol.* *7*, 1000173.
65. Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., and Battaglia, F.P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat. Neurosci.* *12*, 919–926.
66. Sutton, R.S., and Barto, A.G. (2018). *Reinforcement Learning, second edition: An Introduction* (MIT Press).
67. Johnson, A., and Redish, A.D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Netw.* 10.1016/j.neunet.2005.08.009.

68. Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C.R., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* 345, 1616–1620.
69. 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.
70. 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*. 10.1038/nature12475.
71. 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, 1–15.
72. 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. *bioRxiv*, 2020.05.21.108886.
73. Fujisawa, S., and Buzsáki, G. (2011). A 4 Hz oscillation adaptively synchronizes prefrontal, VTA, and hippocampal activities. *Neuron* 72, 153–165.
74. Schultz, W., Dayan, P., and Read Montague, P. (1997). A Neural Substrate of Prediction and Reward.

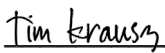
75. Kay, K., and Frank, L.M. (2019). Three brain states in the hippocampus and cortex. *Hippocampus* 29, 184–238.
76. Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., and Mathis, M.W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* 14, 2152–2176.
77. 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.
78. Chung, J.E., Magland, J.F., Barnett, A.H., Tolosa, V.M., Tooker, A.C., Lee, K.Y., Shah, K.G., Felix, S.H., Frank, L.M., and Greengard, L.F. (2017). A Fully Automated Approach to Spike Sorting. *Neuron*. 10.1016/j.neuron.2017.08.030.
79. Kay, K., Sosa, M., Chung, J.E., Karlsson, M.P., Larkin, M.C., and Frank, L.M. (2016). A hippocampal network for spatial coding during immobility and sleep. *Nature* 531, 185–190.
80. Yu, J.Y., Kay, K., Liu, D.F., Grossrubatscher, I., Loback, A., Sosa, M., Chung, J.E., Karlsson, M.P., Larkin, M.C., and Frank, L.M. (2017). Distinct hippocampal-cortical memory representations for experiences associated with movement versus immobility. *Elife*. 10.7554/eLife.27621.
81. Daw, N.D. (2009). Trial-by-trial data analysis using computational models. *Attention & Performance XXIII*.

82. Denovellis, E.L., Frank, L.M., and Eden, U.T. (2019). Characterizing hippocampal replay using hybrid point process state space models. In 2019 53rd Asilomar Conference on Signals, Systems, and Computers (IEEE), pp. 245–249.
83. Joshi, A., Denovellis, E.L., Mankili, A., Meneksedag, Y., Davidson, T.J., Gillespie, A.K., Guidera, J.A., Roumis, D., and Frank, L.M. (2023). Dynamic synchronization between hippocampal representations and stepping. *Nature* 617, 125–131.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

FF74A56E6042448... Author Signature

11/9/2023
Date