From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Neuroscience

2024

Rethinking Prediction Errors in Dopaminergic Signals in the Brain

Lee, Rachel Stephanie

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01qb98mj808

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Rethinking Prediction Errors in Dopaminergic Signals in the Brain

Rachel Stephanie Lee

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE

PRINCETON NEUROSCIENCE INSTITUTE

Adviser: Nathaniel Daw

January 2024

© Copyright by Rachel Stephanie Lee, 2023.
All rights reserved.

Abstract

The hypothesis that midbrain dopamine (DA) neurons broadcast an error for the prediction of reward (reward prediction error, RPE) is among the great successes of computational neuroscience (Houk, Adams, and Barto 1995; Montague, Dayan, and Sejnowski 1996; Schultz, Dayan, and Montague 1997). However, modern empirical results have challenged core aspects of this theory, and my dissertation looks in two such datasets: the first of which delved into how DA responses reflect movement correlates that cannot be explained by a RPE (Parker et al. 2016) and the second of which showed how DA might reflect heterogeneous features rather than convey a scalar, global signal (Engelhard et al. 2019). In my dissertation, I build new models and theories to update the classic RL algorithm TD learning in order to account for these new results. In Chapter 2, I investigate dorsal medial striatum (DMS) projecting DA neurons and how they do not reflect RPE with respect to contralateral movement, but contralateral movement directly and the value of the chosen action. In Chapter 3, I introduce a "feature-specific" PE model in order to explain heterogeneous DA responses in the VTA. I argue that the heterogeneity is a reflection of the input state features upstream of the DA neurons, and show how our model can recapitulate how patterns of heterogeneity might arise for reward prediction errors and movement responses. In Chapter 4, I revisit the dataset from Chapter 2 to determine if the movement correlates could be prediction errors with respect to movements rather than reward. While I could not find an action PE with respect to contralateral movement, I am able to show that DMS-projecting DA population at lever presentation does reflect an identity-free action PE, or an "action-surprise" signal. Overall, my dissertation aims to rethink the PE signal

in TD learning, showing that classic RPE accounts of DA need to be updated to account for these new, puzzling data. Together, my models show that DA data may be better understood as a more generalized family of PE models, reflecting a richer signal than just a scalar RPE depending on the inputs they receive.

Acknowledgements

This thesis is the culmination of my collaborations with wonderful people that I have been so lucky to work with these past few years and the support of family and friends who have been with me through good times and tough times.

First and foremost, my two advisors: Nathaniel Daw and Ilana Witten. Working with the two of you has been a tremendous privilege, and it's been a gift and joy learning from the two of you.

My thesis committee: Ken Norman, whose talk I attended as a sophomore at Stanford made it clear that PNI was where I wanted to be and where I could do the work I could do even before I was sure what type of science I wanted to pursue. Thank you for always being available to provide sage advice.

My collaborators and mentors, past and present: Jay McClelland and Andrew Saxe, who found me when I was an undergraduate at Stanford and were willing to invest time and funding to convince me I could be a scientist. Josh Tenenbaum and Sam Gershman, with whom I've been so lucky to have spent a summer as an undergraduate to work on Bayesian modeling. Jeffrey Chen, for always reminding me what's most important in life and helping me orient to my True North. Nathan Parker, Marcelo Mattar, Yotam Sagiv, and Ben Engelhard, who have been cheerleaders, supporters, and wonderful people to write papers with. Will Dabney, Khimya

Khetarpal, and Kevin Miller, who patiently helped navigate ML research with me during my internship at DeepMind.

To my past and present labmates: In the Daw Lab: Mostafa Abdou, Sean Allan, Yoel Sanchez Araujo, Aaron Bornstein, Flora Bouchacourt, Carlos Correa, Ashley Chung, Neil Garrett, Lindsay Hunter, Ari Kahn, Gili Karni, Ji-sung Kim, Kevin Lloyd, Ida Momennejad, Kate, Nussenbaum, Payam Piray, Dylan Rich, Harrison Ritz, Evan Russek, Sarah Jo Venditto, Oliver Vikbladh, and Sam Zorowitz. In the Witten lab: Akhil Bandi, Scott Bolkan, Lindsey Brown, Lili Cai, Courtney Cameron, Ryan Cho, Claire Choi, Victoria Corbit, Julia Cox, Rob Fetcho, Weston Fleming, Laura Haetzel, Cameron Hayes, Heejae Jang, Emma Keppler, Brenna McMannon, Addie Minerva, Malavika Murugan, Alex Pan Vazquez, Michael Siniscalchi, Iris Stone, Bichan Wu, Lindsay Willmore, Anna Zhukovskaya, and Chris Zimmerman.

The many admin and support staff at PNI: Thank you especially to Ed Clayton, Alex Michaud, and the many other admins and staff whose job helped us both survive and thrive our PhD's. Jawaan Lowery, Sara Lovelace, Gene Piccola, Randee Tengi, and the rest of the PNI IT who have walked me through all of my tricky set ups. Alvaro Luna and Juan Lopez, who helped me investigate quirks in the virtual reality codebase no other experimentalist would encounter.

To my friends within and outside of research: My PNI cohort, who manage to be equal measures brilliant and supportive of each other: Paula Brooks, Kevin Chen, Tomohito Minakuchi, Shruthi Ravindranath, Elizabeth Tawa, Sarah Jo Venditto, Lindsay Willmore, David Zoltowski, Sam Zorowitz. My friends at Princeton: Dakota Blackman, Sally Cochrane,

Sev Harootonian, Marissa Hansen, Neggin Keshavarzian, Bertilia Trieu, Dan Mirea, Olga Lositsky, Orren Karnoil-Tambour, Sebastian Mussilik, Edna Normand. My friends on the West Coast: Eric Yurko, Zeyu Zheng, Erik Burton, Robin Miller.

To my family: Thank you for investing in my education, for dealing with how gosh darn stubborn I am, and for forgiving me even when I am difficult. My parents, Bowei and Christine, who might not have quite understood my fascination with research but always willing to be my 後盾. My wonderful generous siblings, Hendrick and Michelle, who are always willing to feed me a good meal, give me some good advice, and a pat on the back to get me back to work. My extended family on the East Coast: Wayzen and Lauren; Nancy and Jeff; Jamie and Elan; Albert and Stephanie; Sally and Mina 姑姑.

And most of all, thank you to Aaron Jaffey and Joel Cruz, who carried the heaviest of burdens when things got tough. Thank you for holding my hand and making me feel whole and okay even when I felt broken. Thank you for cooking me meals, traveling with me, giving me pep talks, generally loving me as a person, and most of all, humoring me when I melodramatically ask you to pledge that you'd finish my thesis for me if I ever meet my untimely demise.

Table of Contents

Abstract **3**

Acknowledgements **5**

Table of Contents **8**

1. Introduction **10**

1.3 TD Learning as a Dopaminergic Reinforcement Learning Model 15

1.4 Other Dopaminergic Reinforcement Learning Models 17

1.5 Aims of dissertation and overview of chapters 20

2. Reward Prediction Error Does Not Explain Movement Selectivity in DMS-Projecting Dopamine Neurons **21**

2.1 Introduction 22

2.2 Methods 25

2.3 Results 31

2.4 Discussion 43

2.5 Supplementary Materials 50

2.6 Acknowledgements 59

3. A Feature-Specific Prediction Error Model Explains Dopaminergic Heterogeneity **60**

3.1 Introduction 61

3.2 Methods 66

3.3 Results 82

3.4 Discussion 111

3.5 Supplementary Materials 121

3.6 Acknowledgements 127

4. Understanding Value-free Action Prediction Error in the Brain **128**

4.1 Introduction 129

4.2 Methods 131

4.3 Results 133

4.4 Discussion 140

5. Conclusion **143**

5.1 Summary of Results 144

5.2 Revisiting Parker et al. Dataset 144

5.3 Re-evaluating TD learning and RPE signals 146

Bibliography **148**

Chapter 1

Introduction

Reinforcement learning (RL) describes the process of an agent learning to make decisions by trial and error through reward or punishment. RL models are commonly linked to the phasic firing of **Dopamine** (DA) neurons in the midbrain, which is thought to primarily and uniformly encode **reward prediction error** (RPE) (A. G. Barto, Sutton, and Anderson 1983; Schultz, Dayan, and Montague 1997). In the past two decades, establishing this link between DAergic signals and RL models has been a great achievement for the field of computational neuroscience, validated by experimental work spanning calcium imaging, optogenetics, behavior, and electrophysiology (Parker et al. 2016; Chang et al. 2017; Schultz, Apicella, and Ljungberg 1993). Classic neuroscience RL models, such as the temporal difference (TD) learning algorithm, hypothesize that RPE is used as a learning signal to bootstrap the value of various actions without a supervised signal (Sutton 1988). Given some feature state (an input upon which reward expectations and choices are made, summarizes the current and past task situation, and keeps track of the current description of the environment to the agent), the RPE signal updates the weight that maps input state features to values. Empirical evidence has shown how each of these components can be mapped onto the basal ganglia: cortical inputs reflecting state (Y. K. Takahashi et al. 2011; Wilson et al. 2014; Bradfield and Hart 2020), striatal areas reflecting value (Lau and Glimcher 2008; H. Kim et al. 2009; Samejima et al. 2005), and dopamine neurons conveying a single, scalar RPE (Chang et al. 2017; Schultz, Apicella, and Ljungberg 1993; Bayer and Glimcher 2005) .

My thesis challenges some of the core assumptions of TD learning, with each chapter rethinking how PE is traditionally thought of in the brain by modeling and reinterpreting puzzling and interesting DA data. Overall, the goal is to refine and update these classic neuroscience RL

models as they are confronted with empirical data that can no longer be so easily explained by them. Ultimately, understanding how these signals might arise helps us better understand dopaminergic function, and sheds light on crucial questions relating to dopamine's many important functions in the brain. 1.1 Dopamine and its Functions

Dopamine is a major neurotransmitter in the brain, playing an important role in both reward motivated behavior and movement (Niv et al. 2007; Berke 2018). While dopamine is widely transmitted throughout the brain, dopaminergic neurons are highly concentrated in the basal ganglia (Björklund and Dunnett 2007). The primary input nucleus of the basal ganglia is the striatum, which is a part of multiple parallel cortico-subcortical loops and receives dense innervations from midbrain DA neurons as well (Beckstead, Domesick, and Nauta 1979; Swanson 1982; Beier et al. 2015; Poulin et al. 2018). These loops are formed because the inputs to the striatum are topographically organized, and these segregated projections persist within the basal ganglia as well (Alexander and Crutcher 1990; Alexander, DeLong, and Strick 1986). These parallel loops are traditionally thought of to relate to different functions, each forming a motor circuit, an oculomotor circuit, a limbic circuit, and an associative circuit.

Originally, dopamine has been pinpointed as a conduit for motor control, especially since neurodegenerative diseases such as Parkinson's show how the depletion of dopamine leads to the loss of motor function altogether (Kish, Shannak, and Hornykiewicz 1988; Brück et al. 2006). The role of dopamine in movement has been further delineated through direct and indirect pathways in the basal ganglia, showing how D1 versus D2 DA receptors regulate the start and stop of movements (Albin, Young, and Penney 1989; DeLong 1990). Classically, this model

(commonly known as the go/nogo model) argues that the pathways oppositely modulate the firing rates of basal ganglia output nuclei: that is, the direct pathway disinhibits brainstem motor structure to promote movement while the indirect pathway activates basal ganglia output nuclei to suppress their motor targets (Frank 2005).

Yet, single cell recordings of DA conducted in the 90's showed a puzzling result– when DA neurons in monkeys were recorded while pairing auditory cues to rewarding juice drops, the DA neurons did not fire with respect to movement, but to the reward and the conditioned auditory stimuli instead (Schultz, Apicella, and Ljungberg 1993). Specifically, they showed that the DA activity reflected reward prediction error–a teaching signal used in RL models related to the difference between the expected and actually received reward– (Schultz, Apicella, and Ljungberg 1993), and work in the decades since expanded this research to show how DA is crucial for reward motivated behavior and reinforcement learning (Hollerman and Schultz 1998; Schultz 1998; Cohen et al. 2012; Eshel et al. 2016, 2015). In fact, DA neurons not only reflect RPE; stimulation of DA neurons in the basal ganglia can replace RPE signals and help facilitate learning directly (Saunders et al. 2018). While these two functions of DA have been well-established, how DA neurons perform both these functions concurrently remains murky. Understandably, reward-motivated behavior and movement are very much intertwined; however, most theories and0 models have focused on one or the other function, and only a few reconcile how DA can do both (Frank, Seeberger, and O'reilly 2004; Niv et al. 2007; McClure, Daw, and Montague 2003).

The separation between movement and reward-motivated behavior can be reflected in the basal ganglia topography, specifically in dorsal versus ventral pathways (Saunders et al. 2018; Parker et al. 2016). Generally, the ventral tegmental area (VTA) projects to the ventral striatum in a gradient to the SNc, which preferentially projects to the dorsal striatum (Beckstead, Domesick, and Nauta 1979; Swanson 1982; Lerner et al. 2015). This division also reflects the motor vs reward-motivated functions, with VTA DA neurons reflecting RPE and SNc DA neurons reflecting movement activity (Saunders et al. 2018; Parker et al. 2016).

Movement and RPE may be the most prominent functions of DA, but DA is also involved in other related behaviors. Substantia Nigra (SNc) DA axon activity has been shown to be correlated with saliency, for example, activating when unexpected positive and negative events occur (Lerner et al. 2015). Lateral SnC DA neurons (specifically, tail-striatum (TS) projecting DA neurons) have been responsive to threatening stimuli such as loud tones or air puffs (Menegas et al. 2017, 2018).

Recent empirical data have shown how DA can have more heterogeneous responses within projections as well. VTA recordings during an evidence accumulation task has uncovered DA neurons that respond to various reward and non-reward related behavioral features, including the kinematics of the animal, the accuracy of the trial, and the previous trial's reward (Engelhard et al. 2019). VTA DA neurons have also shown how they could reflect optimistically and pessimistically to reward amounts, suggesting that, instead of coding RPE to estimate the median reward, they may be coding the entire distribution of possible rewards (Dabney et al. 2020). Finally, recent work has shown how DA neurons may respond to value-free sensory

changes of the cues: that is, in addition to responding to RPE, they can also signal a change in the juice flavor that the animal is receiving (Y. K. Takahashi et al. 2017). These results, while puzzling, may be orthogonal and all co-occur at the same time. Yet, it's clear that classic TD learning models cannot account for these results, and important work lies ahead of us to update these models to better understand DA functions.

1.3 TD Learning as a Dopaminergic Reinforcement Learning Model

Among biological models of RL, the classic TD learning model (Sutton 1988) is typically used to link DA and RPE (Schultz, Dayan, and Montague 1997; Montague, Dayan, and Sejnowski 1996; Houk, Adams, and Barto 1995). Specifically, the RPE signal at time t is given as

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \quad [1.1]$$

with r_t the reward, $V(s_t)$ the estimated value for a given state s_t, and γ the discount factor (Sutton 1988). This learning signal is used as a bootstrap method to learn the values of different states, given that the value function for each state is defined recursively through the Bellman equation (Bellman 1957):

$$V(s_t) = r_t + E[V(s_{t+1})|s_t] \quad [1.2]$$

This definition essentially argues that the value for a given state would be given by the reward at that state plus the expected values of all subsequent states. Because the TD error includes the future reward expectations given as $V(s_{t+1}) - V(s_t)$, the signal is able to not only learn about the reward itself but also stimuli that are paired with reward. This particular feature of the model helps explain the aforementioned empirical results that showed how DA not only

reflected reward, but the transfer of that reward signal to conditioned cues (Schultz, Apicella, and Ljungberg 1993).

That this TD error signal is scalar is a core aspect of these models. Computationally, this reflects the neuroeconomic concept of common currency—that is, different types of reward in the brain must at some stage be weighed along a single scale in order to allow the system to compare disparate outcomes with the same scale. Since the reward is scalar, value (predictions of future reward that RL models learn to make choices) is scalar, and therefore prediction errors for learning these values (the signal associated with DA) are also scalar.

However successful the TD learning model has been, as empirical methods to recording DA improved and RL tasks became more complex, recent results have challenged whether or not these models are sufficient. For one, the aforementioned results of heterogeneous VTA DA neurons cannot fit cleanly into the TD learning model's scalar RPE signal (Engelhard et al. 2019; Dabney et al. 2020; Y. K. Takahashi et al. 2017). Moreover, the movement related signaling that has been long associated with DA also cannot be clearly addressed by these models.

Another key deficiency to the classic TD learning algorithm is the fact that we know that humans and animals actually interact with complex, high-dimensional environments, which can be difficult or impossible to parse with models only using scalar learning signals. For instance, modern deep network algorithms used to solve these problems in AI start with a single, scalar prediction error, but then break it up into many individual errors tied to each individual weight

in the network, using the (biologically unrealistic) backpropagation algorithm (Mnih et al. 2015, 2016).

In addition, current RL models also have the issue of how state—an important variable in RL models that summarizes the task situation and also accounts for previous instances of the task—can be represented in the brain. Current RL models in neuroscience focus on simple stimuli and tasks, allowing for the models to pick simple state representations to encode the task situations. However, RL models in real-world environments have high-dimensional state spaces, and it remains unclear how the brain might represent the state for the task.

1.4 Other Dopaminergic Reinforcement Learning Models

In order to address DA's heterogeneous responses, various models have emerged that essentially take the classic TD learning model, but replicate the model to have PEs to account for various reward-related DA functions. I have termed this class of models as outcome-specific PE models, as they replace the reward term driving each of the TD learning models to different outcome contingencies, including different aspects of reward or other quantities altogether (Daw, Kakade, and Dayan 2002; Lloyd and Dayan 2016; Gardner, Schoenbaum, and Gershman 2018; Dabney et al. 2020; Akiti et al. 2022; Rao 2010; Lak et al. 2017; Daw, Courville, and Touretzky 2003; Kurth-Nelson and Redish 2009; Gershman, Pesaran, and Daw 2009). Prominent models of this class include distributional RL, which has multiple PEs for different reward contingencies (Dabney et al. 2020), and successor representation, which has PEs for different sensory input features (Gardner, Schoenbaum, and Gershman 2018). Generally, while these models can

account for different DA functions, they are a type of labeled-line model that needs more and more duplicated TD learning schematics for each new feature DA neurons might respond to. They do not altogether show how these new functionalities might emerge, just that they can add more complexities to account for them.

Another class of biological RL models account for two complementary types of reward-motivated behaviors: goal-directed versus habitual control (Balleine and O'Doherty 2010; Dickinson 1985). In particular, goal-directed behaviors describe choosing actions based on the outcomes they might bring about (essentially learning "action-outcome" representations). In contrast, habitual behaviors encompass associating cues in the environment with actions the animal might take (learning some "stimulus-response" relationship, essentially). Animals and agents might switch between these two behaviors, depending on circumstances in the environment (Wood and Neal 2007; Dickinson 1985; Wood and Rünger 2016).

Extensive empirical studies have shown that there are dissociable habitual versus goal directed behavior pathways in the striatum. Specifically, OFC and dorsal medial striatum (DMS) have been associated with goal-directed behaviors (Yin et al. 2005; Jones et al. 2012; K. J. Miller, Botvinick, and Brody 2017; Gremel and Costa 2013) while dorsal lateral striatum (DLS) activity regulates habitual behavior (Yin, Knowlton, and Balleine 2004; Yin and Knowlton 2006; Gremel and Costa 2013; Tricomi, Balleine, and O'Doherty 2009). These distinctions have been shown both in observing neural activity (Gremel and Costa 2013; Tricomi, Balleine, and O'Doherty 2009), and in perturbation and lesioning studies that show how disabling individual pathways

shut down goal-directed and habitual behavior (Yin et al. 2005; Jones et al. 2012; K. J. Miller, Botvinick, and Brody 2017; Yin, Knowlton, and Balleine 2004; Yin and Knowlton 2006).

Generally, computational models have addressed these two behaviors by using "model-based" versus "model-free" controllers (Daw, Niv, and Dayan 2005; Dolan and Dayan 2013; Sutton, Barto, and Others 1998). That is, when engaging in goal-directed behavior, the animal would choose a "model-based" controller that keeps track of action-outcome representations to choose the right actions in the right context to generate the desired outcome. On the other hand, animals would simply use learned associations with a "model-free" controller when engaging in habitual behavior. Given DA's well-documented functions in learning reward-motivated associations, "model-free" controllers are typically considered regulated by DA in the brain (Bayer and Glimcher 2005; Montague, Dayan, and Sejnowski 1996). However, there has been far less evidence showing that the brain does have distinct model-based versus model-free pathways. Regions that respond to "model-based" PE errors also have shown "model-free" correlates (Bornstein and Daw 2011; Doll, Simon, and Daw 2012; D. Lee, Seo, and Jung 2012; Shohamy 2011). DAnergic midbrain that is classically associated with "model-free" algorithms can also show "model-based" analogues (Kishida et al. 2016; Daw et al. 2011). Thus, there remains open questions with respect to how exactly DA might regulate these behaviors, and what models might be the apt way to differentiate between these behaviors.

1.5 Aims of dissertation and overview of chapters

The goal of my thesis is to present new theoretical models to better explain recent empirical findings that challenge the classic TD learning narrative. Each chapter focuses on a different way I can reformulate the PE signal for DA. In Chapter 2, I focus on movement signals in DMS-projecting DA neurons, showing whether or not those movement correlates are motor or value modulated. In Chapter 3, I challenge TD learning algorithm's scalar RPE, showing how a vectorized feature-specific RPE signal can account for heterogeneous DA signals in the VTA during a high-dimensional evidence accumulation task. In Chapter 4, I return to the movement versus reward-motivated functions of DA, seeing if I can find an analogous movement-related PE signal as well as an RPE in DLS-projecting DA neurons, and how that might relate to habitual and goal-directed behavior. All in all, I re-analyze two datasets from the Witten lab. For Chapter 2 and 4, I look at bulk-photometry of DA with different striatal projections while mice perform a reversal learning task. For Chapter 3, I analyze two-photon VTA recordings of mice in a VR evidence accumulation task. Finally, I discuss in the conclusion how these various interpretations of overlapping DA data can be understood altogether, and help us rethink how PEs in DA signals can function.

Chapter 2

Reward Prediction Error Does Not Explain Movement Selectivity in DMS-Projecting Dopamine Neurons

The contents of this chapter have been published as:

Lee, Rachel S., Marcelo G. Mattar, Nathan F. Parker, Ilana B. Witten, and Nathaniel D. Daw. 2019. "Reward Prediction Error Does Not Explain Movement Selectivity in DMS-Projecting Dopamine Neurons." *eLife* 8 (April). https://doi.org/10.7554/eLife.42992.

2.1 Introduction

A central feature of dopamine (DA) is its association with two apparently distinct functions: reward and movement (Niv et al. 2007; Berke 2018). Although manipulation of DA produces gross effects on movement initiation and invigoration, physiological recordings of DA neurons have historically shown few neural correlates of motor events (Wise 2004; Schultz, Dayan, and Montague 1997). Instead, classic studies reported responses to rewards and reward-predicting cues, with a pattern suggesting that DA neurons carry a "reward prediction error" (RPE) – the difference between expected reward and observed reward – for learning to anticipate rewards (Schultz, Dayan, and Montague 1997; Andrew G. Barto 1995; Cohen et al. 2012; Coddington and Dudman 2018; Soares, Atallah, and Paton 2016; Hart et al. 2014). In this classic framework, rather than explicitly encoding movement, DA neurons influence movements indirectly, by determining which movements are learned, and/or the general motivation to engage in a movement (Niv et al. 2007; A. G. E. Collins and Frank 2014; Berke 2018).

Complicating this classic view, however, several recent studies have suggested that subpopulations of DA neurons may have a more direct role in encoding movement (Parker et al. 2016). For example, we recently reported that whereas dopamine neurons projecting to ventral striatum showed classic RPE signals, a subset of midbrain DA neurons that project to the dorsomedial striatum (DMS) were selective for a mouse's choice of action (Parker et al. 2016). In particular, they responded more strongly during contralateral (versus ipsilateral) choices as mice perform a probabilistic learning task (Parker et al. 2016). In addition, there have been several other recent studies that reported phasic changes in DA activity at the onset of

spontaneous movements (Dodson et al. 2016; Howe and Dombeck 2016; da Silva et al. 2018; Barter et al. 2015; Syed et al. 2016). In addition, other studies have shown that DA neurons may also have other forms of apparently non-RPE signals, such as signals related to novel or aversive stimuli (Menegas et al. 2017; Horvitz 2000; Ungless, Magill, and Bolam 2004; M. Matsumoto and Hikosaka 2009; Lammel et al. 2011).

These recent observations of movement selectivity leave open an important question: can the putatively movement-related signals be reconciled with Reinforcement Learning (RL) models describing the classic RPE signal? For instance, while it seems plausible that movement-related DA signals could influence movement via directly modulating striatal medium spiny neurons (DeLong 1990), these signals are accompanied in the same recordings by RPEs which are thought to drive corticostriatal plasticity (Reynolds, Hyland, and Wickens 2001). It is unclear how these two qualitatively different messages could be teased apart by the recipient neurons. Here we introduce and test one possible answer to this question, which we argue is left open by Parker et al.'s (2016) results and also by other reports of movement-related DA activity: that these movement-related signals actually also reflect RPEs, but for reward predictions tied to particular movement direction. Specifically, computational models like the actor-critic (A. G. Barto, Sutton, and Anderson 1983) and advantage learning (Baird 1994) learn separate predictions about the overall value of situations or stimuli and about the value of specific actions. It has long been suggested these two calculations might be localized to ventral vs dorsal striatum, respectively (Montague, Dayan, and Sejnowski 1996; O'Doherty et al. 2004; Y. Takahashi, Schoenbaum, and Niv 2008). Furthermore, a human neuroimaging experiment

reported evidence of distinct prediction errors for right and left movements in the corresponding contralateral striatum (Gershman, Pesaran, and Daw 2009).

This leads to the specific hypothesis that putative movement-related signals in DMS-projecting DA neurons might actually reflect an RPE related to the predicted value of contralateral choices. If so, this would unify two seemingly distinct messages observed in DA activity. Importantly, a choice-specific RPE could explain choice-related correlates observed prior to the time of reward. This is because temporal difference RPEs do not just signal error when a reward is received, they also have a phasic anticipatory component triggered by predictive cues indicating the availability and timing of future reward, such as (in choice tasks) the presentation of levers or choice targets (Montague, Dayan, and Sejnowski 1996; Morris et al. 2006; Roesch, Calu, and Schoenbaum 2007). This anticipatory prediction error is proportional to the value of the future expected reward following a given choice – indeed, we henceforth refer to this component of the RPE as a "value" signal, which tracks the reward expected for a choice. Crucially, a choice-specific value signal can masquerade as a choice signal because, by definition, action and value are closely related to each other: animals are more likely to choose actions they predict have high value. In this case, a value signal (RPE) for the contralateral choice will tend to be larger when that action is chosen than when it is not (Samuelson 1938). Altogether, given the fundamental correlation between actions and predicted value, a careful examination of the neural representation of both quantities, and a clear understanding of if and how they can be differentiated, is required to determine whether or not movement direction signals can be better explained as value-related.

Thus, we examined whether dopamine signals in DMS-projecting DA neurons are better understood as a contralateral movement signal or as a contralateral RPE. To tease apart these two possibilities, we measured neural correlates of value and lateralized movement in our DA recordings from mice performing a probabilistic learning task. Since value predictions are subjective, we estimated value in two ways: 1) by using reward on the previous trial as a simple, theory-neutral proxy, and 2) by fitting the behavioral data with a more elaborate trial-by-trial Q-learning model. We compared the observed DA modulations to predictions based on modulation either by movement direction, and/or the expected value (anticipatory RPE) of contralateral or chosen actions.

Ultimately, our results show that DMS-projecting DA neurons' signals are indeed modulated by value (RPE), but, crucially, this modulation reflected the value of the chosen action rather than the contralateral one. Thus, the value aspects of the signals (which were not lateralized) could not explain the contralateral choice selectivity in these neurons, implying that this choice-dependent modulation indeed reflects modulation by contralateral movements and not value.

2.2 Methods

Mice and Surgeries

This article reports new analysis on data originally reported by (Parker et al. 2016). We briefly summarize the methods from that study here. This article reports on data from 17 male mice

expressing Cre recombinase in the tyrosine hydroxylase promoter ($Th^{\text{IRES-Cre}}$), from which GCaMP6f recordings were obtained from DA neurons via fiber photometry.

In the case of DA terminal recordings, Cre-dependent GCaMP6f virus (AAV5-CAG-Flex-GCamp6f-WPRE-SV40; UPenn virus core, injected titer of 3.53 × 1012 pp per ml) was injected into the VTA/SNc, and fibers were placed in the DMS (M–L ± 1.5, A–P 0.74 and D–V −2.4 mm), with one recording area per mouse (n = 12 recording sites). The recording hemisphere was counterbalanced across mice. The mice were recorded bilaterally, with the second site in nucleus accumbens, which is not analyzed in this paper.

In the case of VTA/SN::DMS cell body recordings, Cre-dependent GCaMP6f virus (AAV5-CAG-Flex-GCamp6f-WPRE-SV40; UPenn virus core, injected titer of 3.53 × 1012 pp per ml) was injected into the DMS, and fibers were placed on the cell bodies in VTA/SNc (M–L ± 1.4, A–P 0.74, D–V −2.6 mm), enabling recordings from retrogradely labeled cells (n=4 mice). Three of the mice were recorded from both hemispheres, providing a total of n = 7 recording sites.

One mouse was used for the GFP recordings as a control condition for VTA/SNc::DMS terminals recordings (**Figure 2.1e**).

Instrumental Reversal Learning Task

The recordings were obtained while the mice performed a reversal learning task in an operant chamber with a central nose poke, retractable levers on each side of the nose poke, and reward delivery in a receptacle beneath the central nose poke.

Each trial began with the illumination of the center nose port. After the mouse entered the nose port, the two levers were presented with a delay that varied between 0-1 seconds. The mouse then had 10 seconds to press a lever, otherwise the trial was classified as an abandoned trial and excluded from analysis (this amounted to <2 % of trials for all mice). After the lever-press, an additional random 0-1 second delay (0.1 second intervals, uniform distribution) preceded either CS- with no reward delivery or CS+ with a 4µl reward of 10% sucrose in H_20. Reward outcomes were accompanied by different auditory stimuli: 0.5 seconds of white noise for CS- and 0.5 seconds of 5 kHz pure tone for CS+. Every trial ended with a constant 3 seconds inter-trial delay.

For the reversal learning, each of the levers either had a high probability for reward (70%) or low probability for reward (10%). Throughout the session, the identity of the high probability lever changed in a pseudorandom schedule; specifically, each block consisted of at least 10 rewarded trials plus a random number of trials drawn from a Geometric distribution of p = 0.4 (mean 2.5). On average, there were 23.23 +/- 7.93 trials per block and 9.67 +/- 3.66 blocks per session. Both reported summary statistics are mean +/- standard deviation.

Data processing

All fiber photometry recordings were acquired at 15 Hz. 2-6 recording sessions were obtained per recording site (1 session/day), and these recordings were concatenated across sessions for all analyses. On average, we had 1307.0 ± 676.01 trials per mouse (858.09 ± 368.56 trials per mouse for VTA/SN::DMS terminals recordings and 448.91 ± 455.61 trials per mouse for VTA/SN::DMS Cell-bodies recordings).

The signals from each recording site were post-processed with a high-pass FIR filter with a passband of 0.375 Hz, stopband of 0.075 Hz, and a stopband attenuation of 10 dB to remove baseline fluorescence and correct drift in baseline. We derived dF/F by dividing the high-pass filtered signal by the mean of the signal before high-pass filtering. We then z-scored dF/F for each recording site, with the mean and standard error calculated for the entire recording from each site.

The VTA/SN::DMS terminals data consisted of 10108 total trials across 12 recording sites, and VTA/SN::DMS cell-bodies consisted of 4938 total trials across 7 recording sites.

Q Learning Mixed Effect Model

We fitted a trial-by-trial Q-learning mixed effect model to the behavioral data from each of the 12 mice on all recording sites, and combined data across mice with a hierarchical model. The model was initialized with a Q value of 0 for each action and updated at each trial according to:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha(r_t - Q_t(c_t))$$

where Q is the value for both options, c_t is the option chosen on trial t (lever either contralateral or ipsilateral to recording site), and $0 <= \alpha <= 1$ is a free learning rate parameter. The subject's probability to choose choice c was then given by a softmax equation:

$$P(c_t = c) \propto exp(\beta \cdot Q_t(c) + stay \cdot I(c, c_{t-1}))$$

where β is a free inverse temperature parameter, *stay* is a free parameter encoding how likely the animal will repeat its choice from the last trial, and I is a binary indicator function for choice repetition (1 if c was chosen on the previous trial; 0 otherwise). The three free parameters of the model were estimated separately for each subject, but jointly (in a hierarchical random effects model) with group-level mean and variance parameters reflecting the distribution, over the population, of each subject-level parameter.

The parameters were estimated using Hamiltonian Monte Carlo, as implemented in the Stan programming language (version 2.17.1.0; (Carpenter et al. 2017)). Samples from the posterior distribution over the parameters were extracted using the Python package PyStan (Carpenter et al. 2017). We ran the model with 4 chains of 1,000 iterations for each (of which the first 250 were discarded for burn-in), and the parameter adapt_delta set to 0.99. We verified convergence by visual inspection and by verifying that the potential scale reduction statistic Rhat (Gelman and Rubin 1992) was close to 1.0 (<0.003 for all parameters) (**Table 2.1**).

We used the sampled parameters to compute per-trial Q values for each action, trial, and mouse. We calculated the difference between the Q values of the chosen action and unchosen action for each trial. We binned the difference of these Q values for each trial and plotted the average GCaMP6f time-locked to lever presentation for each bin (**Figure 2.3b, e**).

Regression Model

In **Figure 3c,f**, we performed a linear mixed effect model regression to predict GCaMP6f signal at each time point based on Q-values, choice (contralateral vs ipsilateral), their interaction, and an intercept. We took the difference of Q values for the chosen vs unchosen levers, then we standardized the difference of Q values for each mouse and each recording site. GCaMP6f was time-locked to lever presentation, regressing to data points 1 second before and 2 seconds after the time-locked event for 45 total regressions. The regression, as well as the calculation of p values, was performed with the MixedModels package in Julia (Bezanson et al. 2014). The p values were corrected for false discovery rate over the ensemble of timepoints for each regression variable separately, using the procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995) via the MultipleTesting package in Julia (Bezanson et al. 2014).

Multiple event Kernel Analysis

In **Figure 2.3.5,** we fitted a linear regression model to determine the contributions to the ongoing GCaMP6f signal of three simultaneously modeled events (Nose poke, lever presentation, lever press). To do this, we used kernels, or sets of regressors covering a series of time lags

covering the period from 1 second before to 2 seconds after each event. Each event had four kernels, corresponding to the four conditions from **Figure 2.3a, c** (all combinations of contralateral vs ipsilateral trials and previous reward vs no previous reward trials). We solved for the kernels by regressing the design matrix against GCaMP6f data using least squares in R with the rms package (Harrell 2018). The standard error (colored fringes) was calculated using rms' robcov (cluster robust-covariance) function to correct for violations of ordinary least squares assumptions due to animal-by-animal clustering in the residuals.

2.3 Results

Task, behavior and DA recordings

Mice were trained on a probabilistic reversal learning task as reported previously (Parker et al. 2016). Each trial began with an illumination in the nose port, which cued the mouse to initiate a nose poke (**Figure 2.1a**). After a 0-1s delay, two levers appeared on both sides of the nose port. Each lever led to reward either with high probability (70%) or low probability (10%), with the identity of the high probability lever swapping after a block of variable length (see Methods for more details, **Figure 2.1b**). After another 0-1s delay, the mouse either received a sucrose reward and an accompanying auditory stimulus (positive conditioned stimulus, or CS+), or no reward and a different auditory stimulus (negative conditioned stimulus, or CS-).

Given that block transitions were not signaled to the mouse, after each transition mice gradually learned to prefer the lever with the higher chance of reward. To capture this learning, we fitted

their choices using a standard trial-by-trial Q-learning model that predicted the probability of the animal's choice at each trial of the task (**Figure 2.1c, Table 2.1**). In the model, these choices were driven by a pair of decision variables (known as Q-values) putatively reflecting the animal's valuation of each option.

As mice performed this task, we recorded activity from either the terminals or cell bodies of DA neurons that project to DMS (VTA/SN::DMS) using fiber photometry to measure the fluorescence of the calcium indicator GCaMP6f (**Figure 2.1d,e; Figure 2.1.1a,b**). As previously reported, this revealed elevated activity during contralateral choice trials relative to ipsilateral choice trials, particularly in relation to the nose poke and lever presentation events (**Figure 2.1f,g; Figure 2.1.1c**) (Parker et al. 2016).

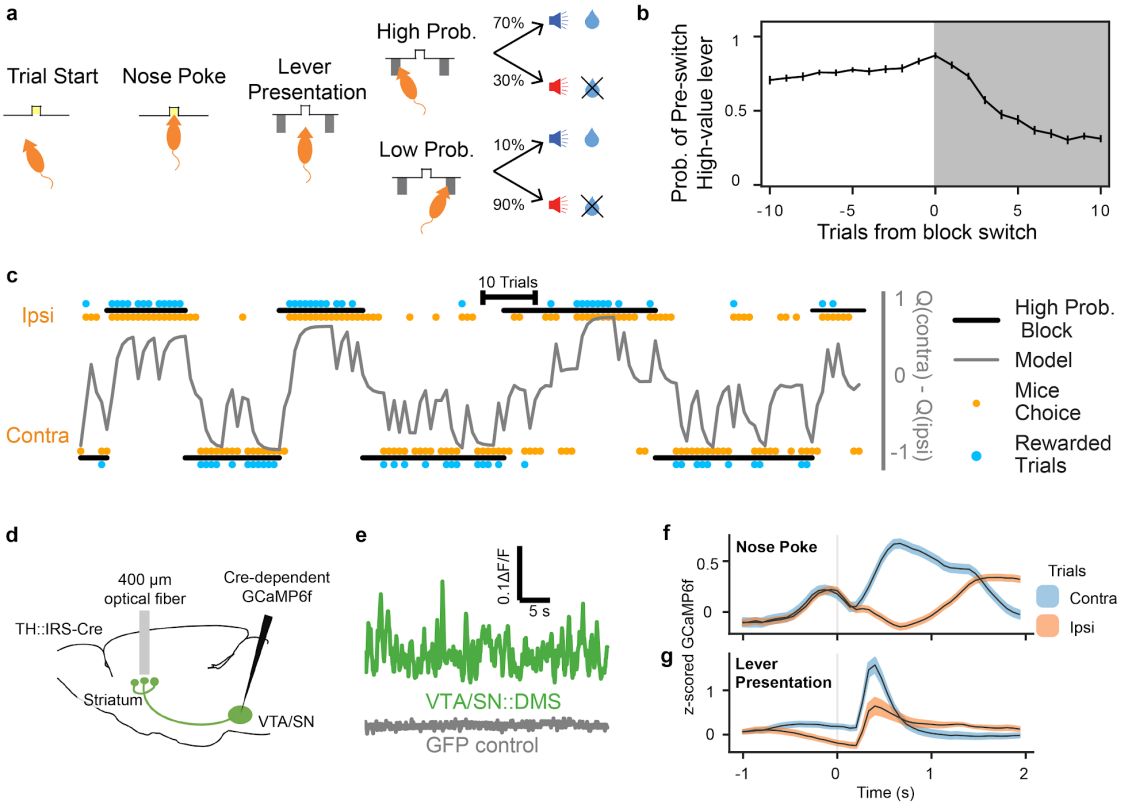


Figure 2.1: Mice performed a probabilistic reversal learning task during GCaMP6f recordings from VTA-SN::DMS terminals or cell bodies. (**a**) Schematic of a mouse performing the task. The illumination of the central nosepoke signaled the start of the trial, allowing the mouse to enter the nose port. After a 0-1 second jitter delay, two levers are presented to the mouse, one of which results in a reward with high probability (70%) and the other with a low probability (10%). The levers swapped probabilities on a pseudorandom schedule, unsignaled to the mouse. (**b**) The averaged probability of how likely the mice were to choose the lever with high value before the switch, 10 trials before and after the block switch, when the identity of the high value lever reversed. Error bars indicate +/- 1 standard error (n = 19 recording sites). "Contra" and "Ipsi" refer to the location of the lever relative to the side of the recording. (**c**) We fitted behavior with a trial-by-trial Q learning mixed effect model. Example trace of 150 trials of a

mouse's behavior compared to the model's results. Black bars above and below the plot indicate which lever had the high probability for reward; Orange dots indicate the mouse's actual choice; Blue dots indicate whether or not mouse was rewarded; Grey line indicate the difference of the model's Q values for contralateral and ipsilateral choices. **d**) Surgical schematic for recording with optical fibers from the GCaMP6f terminals originating from VTA/SN. Projections were determined using viral traces. (**e**) Sample GCaMP6f traces from VTA/SN::DMS terminals and a GFP control animal. **(f, g)** Previous work has reported contralateral choice selectivity in DMS DA terminals (Parker et al. 2016) when the signals are time-locked to nose poke (**f**) and lever presentation (**g**). Colored fringes represent +/- 1 standard error.

	25th Percentile	**50th Percentile (median)**	**75th Percentile**
Alpha (learning rate)	0.205782	0.283441	0.357970
Beta (inverse temperature)	0.990275	1.058405	1.204639
Stay	0.883670	0.945385	1.008465

Table 2.1: **Fitted Parameters for Q-learning model from PyStan.** 25th, 50th, and 75th percentile of the alpha, beta, and stay parameters of the Q-learning mixed effect model. These are the the group-level parameters that reflect the distribution of the subject-level parameters.

Predictions of Contralateral and Chosen Value Models

In order to examine how value-related activity might (or might not) explain seemingly movement-related activity, we introduced two hypothetical frames of reference by which the DMS DA neurons' activity may be modulated by predicted value during trial events prior to the outcome: the DA signals could be modulated by the value of the contralateral option (relative to ipsilateral; **Figure 2.2a**) or by the value of the chosen option (relative to unchosen; **Figure 2.2b**). Note that both of these modulations could be understood as the anticipatory component (occasioned at lever presentation) of a temporal difference RPE, with respect to the respective action's value.

The first possibility is modulation by the value of the contralateral (relative to ipsilateral) action (**Figure 2.2a**; such signals have been reported in human neuroimaging, Gershman et al., 2009, Palmenteri et al. 2009; but not previously to our knowledge examined in DA unit recordings in animals). The motivation for this hypothesis is that, if neurons in DMS participate in contralateral movements, such a side-specific error signal would be appropriate for teaching them when those movements are valuable. In this case, the relative value of the contralateral (versus ipsilateral) choice modulates signals, regardless of whether the choice is contralateral or ipsilateral. Thus, when the DA signals are broken down with respect to both the action chosen and its value, the direction of value modulation would depend on the choice: signals are highest for contralateral choices when these are relatively most valuable, but lowest for ipsilateral choices when *they* are most valuable (because in this case, contralateral choices will be relatively less valuable). Assuming mice tend to choose the option they expect to deliver more reward, such signals would be larger, on average, during contralateral choices than ipsilateral ones

(**Figure 2.2a**), which could in theory explain the contralateral choice selectivity that we observed (**Figure 2.1f,g**).

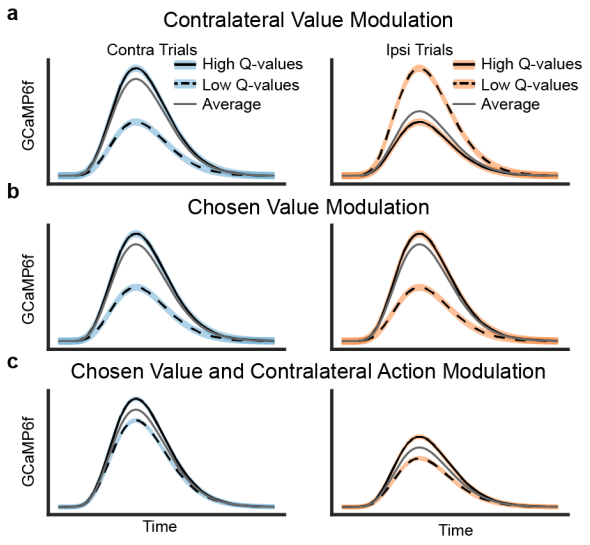


Figure 2.2: Schematics of possible types of value modulation at lever presentation. Trials here are divided based on Q values of chosen minus unchosen action. **(a)** Contralateral value modulation theory postulates that the signals are selective for the *value* of the contralateral action (relative to ipsilateral value) instead of the action itself. This means that the direction of value modulation should be flipped for contralateral versus ipsilateral choices. Since mice would more often choose an option when its value is higher, the average GCaMP6f signals would be higher for contralateral than ipsilateral choices. **(b)** Alternatively, the signals may be modulated by the value of the chosen action, resulting in similar value modulation for contralateral and ipsilateral choice. This type of value modulation will not in itself produce contralateral selectivity seen in previous results. (**c**) However, if the signals were modulated by the chosen value and the contralateral choice, the averaged GCaMP6f would exhibit the previously seen contralateral selectivity.

The second possibility is that value modulation is relative to the chosen (versus unchosen) option (**Figure 2.2b**). This corresponds to the standard type of "critic" RPE most often invoked in models of DA: that is, RPE with respect to the overall value of the current state or situation (where that state reflects any choices previously made), and not specialized to a particular class of action. Indeed, human neuroimaging studies have primarily reported correlates of the value of the chosen option in DAtarget areas (Daw et al., 2006; Boorman et al., 2009; Li & Daw, 2011), and this also has been observed in primate DAneurons (Morris et al., 2006).

If DMS-projecting DA neurons indeed display chosen value modulation (**Figure 2.2b**), rather than contralateral value modulation, the value modulation for both contralateral and ipsilateral choices would be similar. In this case, value modulation could not in itself account for the neurons' elevated activity during contralateral trials, which we have previously observed(**Figure 2.1f,g**). Therefore, to account for contralateral choice preference, one would have to assume DA neurons are also selective for the contralateral action itself (unrelated to their value modulation; **Figure 2.2c**).

DA in dorsomedial striatum is modulated by chosen value, not contralateral value

Next, we determined which type of value modulation better captured the signal in DA neurons that project to DMS by comparing the GCaMP6f signal in these neurons for high and low value trials. We focused on the lever presentation since this event displayed a clear contralateral preference (**Figure 2.1g**). As a simple and objective proxy for the value of each action (i.e., the

component of the RPE at lever presentation for each action), we compared signals when the animal was rewarded (high value), or not (low value), on the previous trial. (To simplify interpretation of this comparison, we only included trials in which the mice made the same choice as the preceding trial, which accounted for 76.6% of the trials.) The traces (**Figure 2.3a**) indicated that the VTA/SN::DMS terminals were modulated by the previous trial's reward. The value-related signals reflected chosen value – responding more when the previous choice was rewarded, whether contralateral or ipsilateral – and therefore did not explain the movement-related effect. This indicates that the DMS-projecting DA neurons represent both chosen value and movement direction (similar to **Figure 2.2c**). The effect of contralateral action modulation was also visible in individual, non-z-scored data in both VTA/SN::DMS terminals (**Figure 2.3.1**) and VTA/SN::DMS cell-bodies (**Figure 2.3.2)**.

We repeated this analysis using trial-by-trial Q values extracted from the model, which we reasoned should reflect a finer grained (though more assumption-laden) estimate of the action's value. (For this analysis, we were able to include both stay and switch trials.) Binning trials by chosen (minus unchosen) value, a similar movement effect and value gradient emerged as we have seen with the previous trial outcome analysis (**Figure 2.3b**). Trials with higher Q values had larger GCaMP6f signals, regardless which side was chosen, again suggesting that VTA/SN::DMS terminals were modulated by the expected value of the chosen (not contralateral) action, in addition to being modulated by contralateral movement.

To quantify these effects statistically, we used a linear mixed effects regression at each of time point of the time-locked GCaMP6f. The explanatory variables included the action chosen

(contra or ipsi), the differential Q values (oriented in the reference frame suggested by the data, chosen minus unchosen), the value by action interaction, and an intercept (**Figure 2.3c**). The results verify significant effects of both movement direction and action value; that is, although a significant value effect is seen, it does not explain away the movement effect. Furthermore, the appearance of a consistent chosen value effect across both ipsilateral and contralateral choices is reflected in a significant value effect and no significant interaction during the period when action and value coding are most prominent (0.25 - 1 seconds after lever presentation), as would have been predicted by the contralateral value model. (There is a small interaction between the variables earlier in the trial, before 0.25 seconds, reflecting small differences in the magnitude of value modulation on contralateral versus ipsilateral trials.) Conversely, when the regression is re-estimated in terms of contralateral value rather than chosen value, a sustained, significant interaction does emerge, providing formal statistical support for the chosen value model; see **Figure 2.3.3.**

We performed the same value modulation analyses on the cell bodies, rather than terminals, of VTA/SN::DMS neurons (**Figure 2.3d-f**). This was motivated by the possibility that there may be changes in neural coding between DA cell bodies and terminals due to direct activation of DA terminals. In this case, we found very similar modulation by both chosen value and contralateral movement in both recording locations.

To verify the robustness of these findings, we conducted further followup analyses. In one set of analyses, we investigated to what extent the DA signals might be tied to particular events other than the lever presentation. First, we repeated our analyses on DA signals time-locked to nose

poke event (**Figure 2.3.4**), and found the same basic pattern of effects. The effect is still clearest close to the average lever presentation latency, suggesting that the modulation of DA signals is more closely related to lever presentation. To more directly verify that our conclusions are independent of the specific choice event alignment, we fitted a linear regression model with kernels capturing the contribution of three three different events (Nose Poke, Lever Presentation, and Lever Press) simultaneously (**Figure 2.3.5**). The results of this multiple event regression were consistent with the simpler single-event regression in **Figure 2.3a, d**.

Next, we examined a few other factors that might have affected movement-specific activity. Taking advantage of the fact that the VTA/SN::DMS cell-bodies data had recordings from both hemispheres in three animals, we directly compared signals across hemispheres in individual mice and observed that the side-specific effects reversed within-animal (**Figure 2.3.6**). This speaks against the possibility that they might reflect animal-specific idiosyncrasies such as side biases. Finally, we considered whether the contralateral action modulation might in part reflect movement vigor rather than action value. We addressed this by repeating the analysis in **Figure 2.3c,f,** but including as an additional covariate the log lever-press latency, as a measure of the action's vigor. For both VTA/SN::DMS terminals and cell-bodies data, the lever-press latency was not a strong predictor for GCaMP6f signals, and the effect of the original predictors largely remained the same (**Figure 2.3.7**).

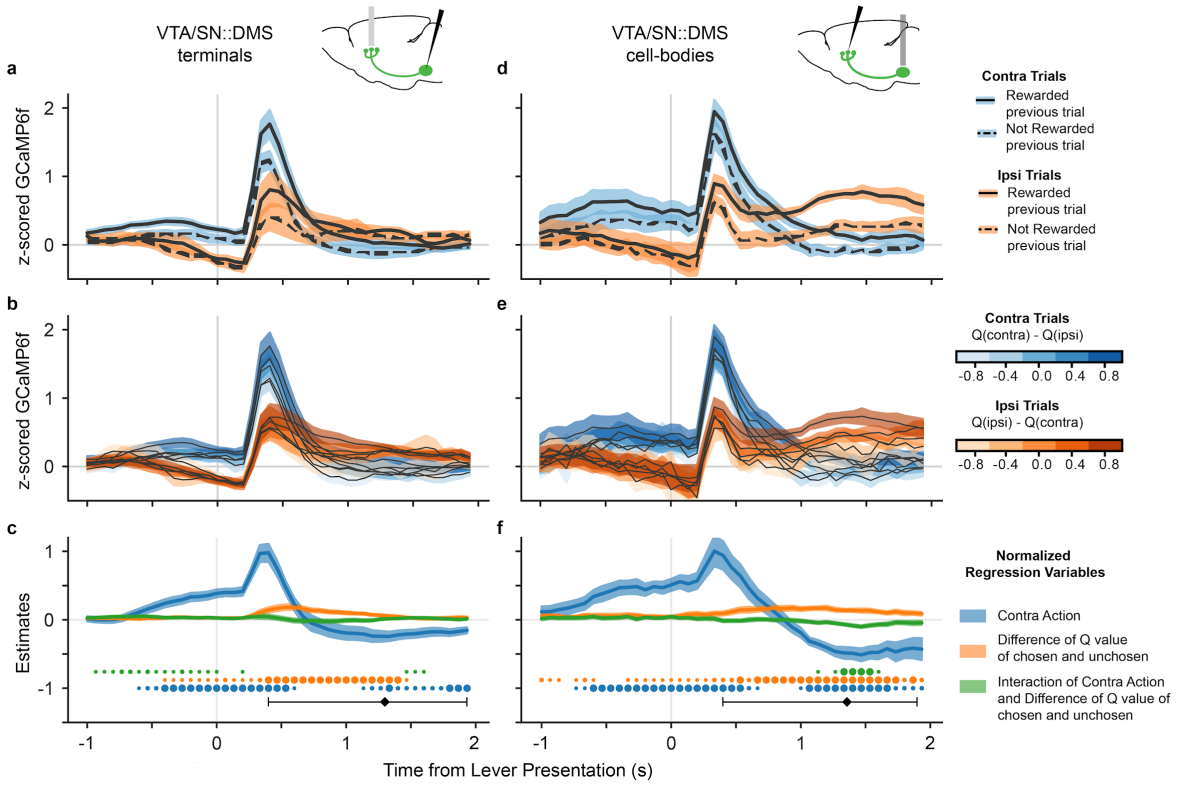


Figure 2.3: DA neurons that project to DMS are modulated by both chosen value and movement direction. (a) GCaMP6f signal time-locked to lever presentation for contralateral trials (blue) and ipsilateral trials (orange), as well as rewarded (solid) and non-rewarded previous trial (dotted) from VTA/SN::DMS terminals. Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n = 12). **(b)** GCaMP6f signal for contralateral trials (blue) and ipsilateral trials (orange), and further binned by the difference of Q values of chosen and unchosen action. Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n = 12). **(c)** Mixed effect model regression on each datapoint from 3 seconds of GCaMP6f traces. Explanatory variables include the action of the mice (blue), the difference in Q values for chosen and unchosen actions (orange), their interaction (green), and an intercept. Colored fringes represent +/- 1 standard error from

estimates (n = 12 recording sites). Black diamond represents the average latency for mice pressing the lever, with the error bars showing the spread of 80% of the latency values. Dots at bottom mark timepoints when the corresponding effect is significantly different from zero at $p<.05$ (small dot), $p<.01$ (medium dot), $p<.001$ (large dot). P values were corrected with Benjamini Hochberg procedure. **(d-f)** Same as **(a-e),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

Direction of movement predicts DMS DA signals

An additional observation supported the interpretation that the contralateral choice selectivity in DMS-projecting DA neurons is related to the direction of movement, and not the value of the choice. When the signals are time-locked to the lever press itself, there is a reversal of the signal selectivity between contralateral and ipsilateral trials, shortly after the lever press (**Figure 2.4**). Although body tracking is not available, this event coincided with a reversal in the animal's physical movement direction, from moving toward the lever from the central nosepoke before the lever press, to moving back to the central reward port after the lever press. In contrast, there is no reversal in the value modulation at the time of the lever press. The fact that the side-specific modulation (and not the value modulation) followed the mice's movement direction during the trial further indicates that movement direction explains the choice selectivity in these DA neurons, and resists explanation in terms of RPE-related signaling.

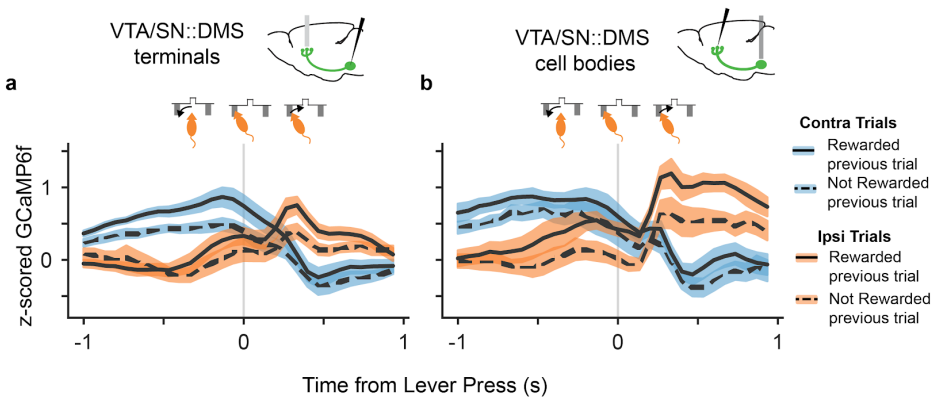


Figure 2.4: DA neurons that project to DMS reverse their choice selectivity after the lever press, around the time the mice reverse their movement direction. (a) GCaMP6f signal from VTA/SN::DMS terminals time-locked to the lever press, for contralateral choice trials (blue) and ipsilateral choice trials (orange), as well as rewarded (solid) and non-rewarded previous trial (dotted). The GCaMP6f traces for each choice crosses shortly after the lever-press, corresponding to the change in the mice's head direction around the time of the lever press (shown schematically above the plot). Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n = 12). **(b)** Same as **(a),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

2.4 Discussion

Recent reports of qualitatively distinct DA signals - movement and RPE-related - have revived perennial puzzles about how the system contributes to both movement and reward, and more specifically raise the question whether there might be a unified computational description of both components in the spirit of the classic RPE models (Parker et al. 2016; Berke 2018;

Coddington and Dudman 2018; Syed et al. 2016). Here we introduce and test one possible route to such a unification: action-specific RPEs, which could explain seemingly action-selective signals as instead reflecting RPE related to the value of those actions. To investigate this possibility, we dissected movement direction and value selectivity in the signals of terminals and cell bodies of DMS-projecting DA neurons (**Figure 2.3**). Contrary to the hypothesis that lateralized movement-related activity might reflect a RPE for contralateral value, multiple lines of evidence clearly indicated that the neurons instead contain distinct movement- and value-related signals, tied to different frames of reference. We did observe value-related signals preceding and following the lever press, which we did not previously analyze in the DMS signal and which are consistent with the anticipatory component of a classic RPE signal (Parker et al. 2016). But because these were modulated by the value of the chosen action, not the contralateral one, they cannot explain the side-specific movement selectivity. The two signals also showed clearly distinct time courses; in particular, the side selectivity reversed polarity following the lever press, but value modulation did not.

Our hypothesis that apparently movement-related DA correlates might instead reflect action-specific RPEs (and our approach to test it by contrasting chosen vs. action-specific value) may also be relevant to other reports of DAergic movement selectivity. For example, Syed et al. recently reported that DA release in the nucleus accumbens (NAcc) was elevated during "go", rather than "no-go", responses, alongside classic RPE-related signals (Syed et al. 2016). This study leaves open a question analogous to the one we raise about Parker's (Parker et al. 2016) DMS results: could NAcc DA instead reflect an RPE specific for "go" actions? This possibility would be consistent with the structure's involvement in appetitive approach and invigoration

(Parkinson et al. 2002), and might unify the RPE- and "go"-related activity reported there via an action-specific RPE (argument analogous to **Figure 2.2a)**. The analyses in the Syed et al. study did not formally compare chosen- vs. action-specific value, and much of the reward-related activity reported there appears consistent with either account (Syed et al. 2016). However, viewed from the perspective of our current work, the key question becomes whether the value-related DA signals on "go" cues reverses for "no-go" cues, as would be predicted for an action-specific RPE. There is at least a hint (albeit significant only at one timepoint in Syed et al.'s Supplemental Figure 9E) that it does not do so (Syed et al. 2016). This suggests that NAcc may also have parallel movement-specific and chosen value signals, which would be broadly confirmatory for our parallel conclusions about DMS-projecting DA neurons.

The RPE account of the DA signal has long held out hope for a unifying perspective on the system's dual roles in movement and reward by proposing that the system's reward-related signals ultimately affect movement indirectly, either by driving learning about movement direction preferences (Montague, Dayan, and Sejnowski 1996) or by modulating motivation to act (Niv et al. 2007). This RPE theory also accounts for multiple seemingly distinct components of the classic DA signal, including anticipatory and reward-related signals, and signals to novel neutral cues. However, the present analyses clearly show that side-specific signals in DMS resist explanation in terms of an extended RPE account, and may instead simply reflect planned or ongoing movements.

Specifically, our results are consistent with the longstanding suggestion that DA signals may be important for directly initiating movement. Such a signal may elicit or execute contralateral

movements via differentially modulating the direct and indirect pathways out of the striatum (Alexander and Crutcher 1990; A. G. E. Collins and Frank 2014; DeLong 1990). The relationship between unilateral DA activity and contralateral movements is also supported by causal manipulations. For instance, classic results demonstrate that unilateral 6-hydroxydopamine (6-OHDA) lesions increase ipsilateral rotations (Costall, Naylor, and Pycock 1976; Ungerstedt and Arbuthnott 1970). Consistent with those results, a recent study reports that unilateral optogenetic excitation of midbrain DA neurons in mice led to contralateral rotations developed over the course of days (Saunders et al. 2018). Importantly, however, our own results are correlational, and we cannot rule out the possibility that the particular activity we study could be related to a range of functions other than movement execution, such as planning or monitoring. Another function that is difficult to distinguish from movement execution is the motivation to move. Although motivation is a broad concept and difficult to operationalize fully, our results address two aspects of it. First, one way to quantify the motivation to act is by the action's predicted value; thus, our main result is to rule out the possibility that neural activity is better accounted for by this motivational variable. We also show that lever press latency (arguably another proxy for motivation) does not explain the DA signals (**Figure 2.3.7**).

Although the movement-related DA signal might be appropriate for execution, it is less clear how it might interact with the plasticity mechanisms hypothesized to be modulated by RPE aspects of the DA signal (Frank, Seeberger, and O'reilly 2004; Steinberg et al. 2013; Reynolds and Wickens 2002). For instance, how would recipient synapses distinguish an RPE component of the signal (appropriate for surprise-modulated learning) from an overlapping component more

relevant to movement elicitation (Berke 2018)? We have ruled out the possibility that the activity is actually a single RPE for action value, but there may still be other sorts of plasticity that might be usefully driven by a purely movement-related signal. One possibility is that plasticity in the dorsal striatum itself follows different rules, which might require an action rather than a prediction error signal (Saunders et al. 2018; Yttri and Dudman 2016). For instance, it has been suggested that some types of instrumental learning are correlational rather than error-driven (Doeller, King, and Burgess 2008) and, more specifically, an early model of instrumental learning ((Guthrie 1935) recently revived by (K. Miller, Shenhav, and Ludvig 2019) posits that stimulus-response habits are not learned from an action's rewarding consequences, as in RPE models, but instead by directly memorizing which actions the organism tends to select in a situation. Although habits are more often linked to adjacent dorsolateral striatum (Yin, Knowlton, and Balleine 2004), a movement signal of the sort described here might be useful to drive this sort of learning. Investigating this suggestion will likely require new experiments centered around causal manipulations of the signal. Overall, our results point to the need for an extended computational account that incorporates the movement direction signals as well as the RPE ones.

Another striking aspect of the results is the co-occurrence of two distinct frames of reference in the signal. Lateralized movement selectivity tracks choices contralateral versus ipsilateral of the recorded hemisphere –appropriate for motor control–, but the value component instead relates to the reward expected for the chosen, versus unchosen, action. This value modulation by the chosen action is suitable for a classic RPE for learning "state" values (since overall value expectancy at any point in time is conditioned on the choices the animal has made; (Morris et

al. 2006), and also consistent with the bulk of BOLD signals in human neuroimaging, where value-related responding throughout dopaminergic targets tends to be organized on chosen-vs-unchosen lines (Daw et al. 2006; Boorman et al. 2009; Li and Daw 2011; O'Doherty 2014).

At the same time, there have been persistent suggestions that given the high dimensionality of an organism's action space, distinct action-specific error signals would be useful for learning about different actions (Russell and Zimdars 2003; Frank and Badre 2012; Diuk et al. 2013) or types of predictions (Gershman and Schoenbaum 2017; Lau, Monteiro, and Paton 2017). Along these lines, there is evidence from BOLD neuroimaging for contralateral error and value signals in the human brain (Gershman, Pesaran, and Daw 2009; Palminteri et al. 2009). Here, we have shown how a similar decomposition might explain movement-related DA signals, and also clarified how this hypothesis can be definitively tested. Although the current study finds no evidence for such laterally decomposed RPEs in DMS, the decomposition of error signals remains an important possibility for future work aimed at understanding heterogeneity of dopamine signals, including other anomalous features like ramps (Howe et al. 2013; Berke 2018; Gershman 2014; Hamid et al. 2016; Engelhard et al. 2018; da Silva et al. 2018). Recent studies, for instance, have shown that midbrain DA neurons may also encode a range of behavioral variables, such as the mice's position, their velocity, their view-angle, and the accuracy of their performance (Howe et al. 2013; da Silva et al. 2018; Engelhard et al. 2018). Our modeling provides a framework for understanding how these DA signals might be interpreted in different reference frames and how they might ultimately encode some form of RPEs with respect to different behavioral variables in the task.

Interestingly, our results were consistent across both recording sites with DMS-projecting DA neurons: the cell bodies and the terminals (**Figure 2.3d-f, Figure 2.4b**). This indicates that the movement selectivity is not introduced in DA neurons at the terminal level, e.g. via striatal cholinergic interneurons or glutamatergic inputs (Kosillo et al. 2016).

An important limitation of the study is the use of fiber photometry, which assesses bulk GCaMP6f signals at the recording site rather than resolving individual neurons. Thus it remains possible that individual neurons do not multiplex the two signals we observe, and that they are instead segregated between distinct populations. Future work should use higher resolution methods to examine these questions at the level of individual DA neurons. A related limitation of this study is the relatively coarse behavioral monitoring; notably, we infer that the reversal in selectivity seen in **Figure 2.4** reflects a change in movement direction, but head tracking would be required to verify this more directly. More generally, future work with finer instrumentation could usefully dissect signal components related to finer-grained movements, and examine how these are related to (or dissociated from) value signals.

2.5 Supplementary Materials

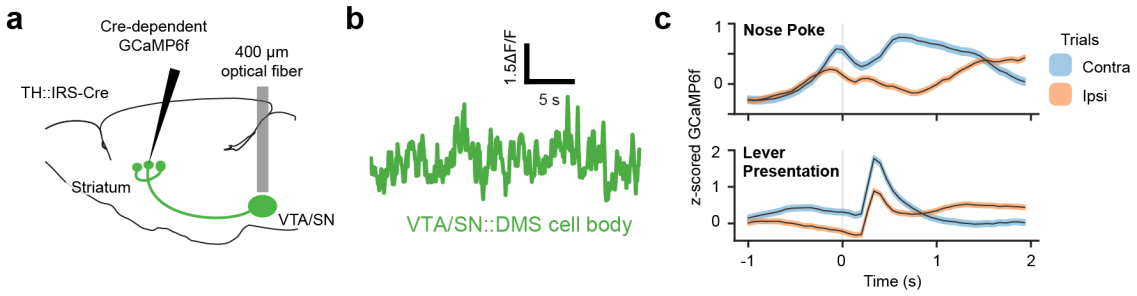


Figure 2.1.1: Recording from VTA/SN::DMS cell bodies (n = 7 recording sites) (**a**) Surgical schematic for recording with optical fibers from the GCaMP6f VTA/SN::DMS cell-bodies. Projections were determined using viral traces. (**b**) Sample GCaMP6f traces from VTA/SN::DMS cell bodies. **(c, d)** Contralateral choice selectivity was also observed in DMS DA cell bodies when the signals were time-locked to nose poke (**c**) and lever presentation (**d**). Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n = 7).

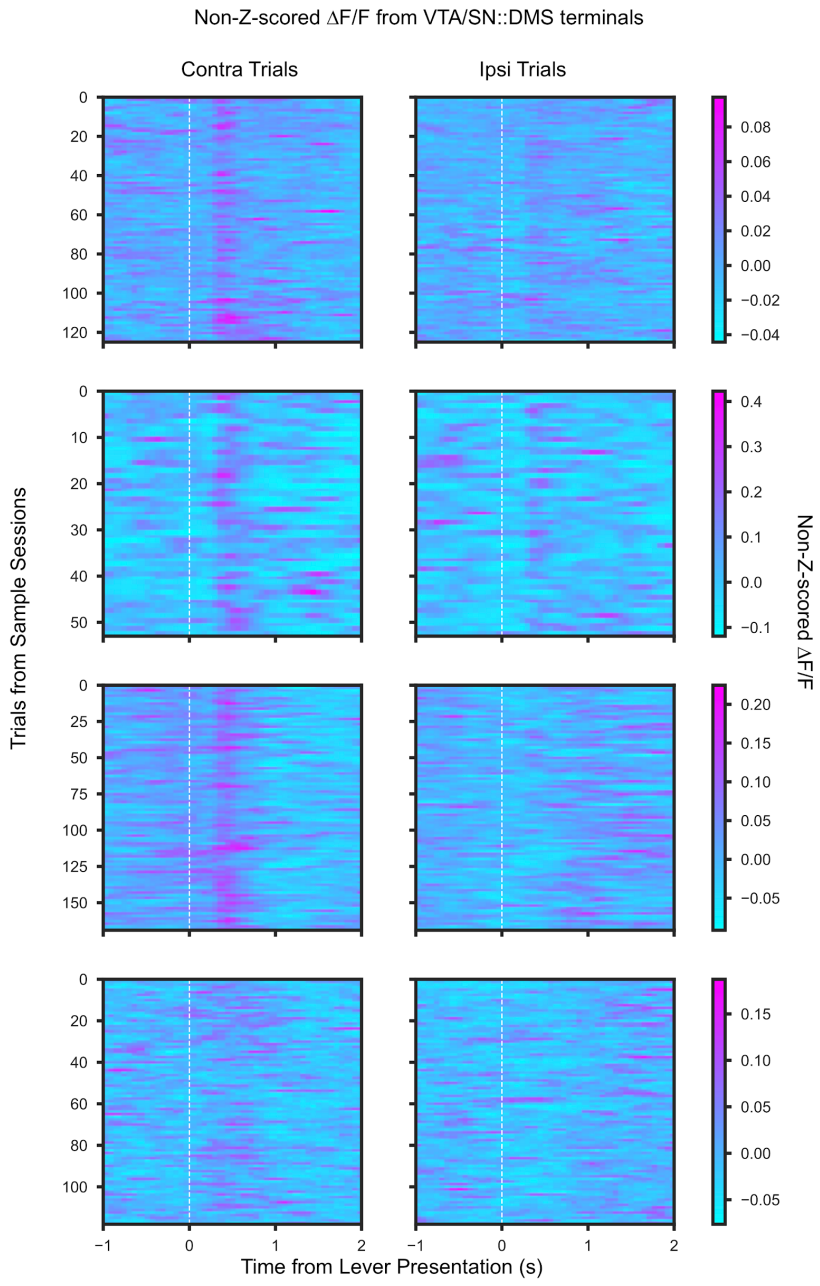
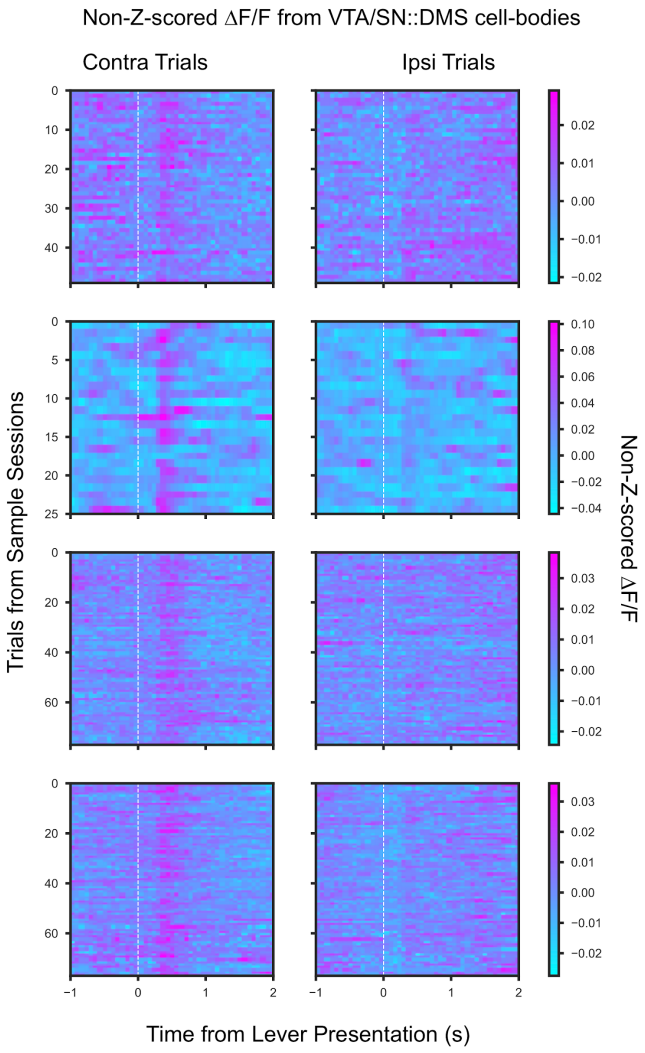


Figure 2.3.1: Four Examples of non-Z-scored Individual Sessions of Photometry Data from VTA/SN::DMS Terminals. Sample, not Z-scored ΔF/F recording from VTA/SN::DMS Terminal. Each row is an example session from a different mouse. Traces are time-locked to the lever presentation for contralateral trials (**left column**) and ipsilateral trials (**right column**).

White dotted vertical line indicate lever presentation. Colorbars are provided for each row for each example session.

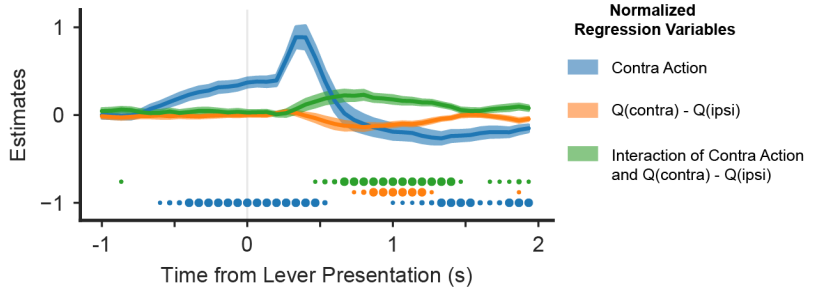
Figure 2.3.2: Four Examples of non-Z-scored Individual Sessions of Photometry Data from VTA/SN::DMS Cell-Bodies. Sample, not Z-scored $\Delta F/F$ recording from VTA/SN::DMS Cell-bodies. Each row is an example session from a different mouse. Traces are time-locked to the lever presentation for contralateral trials (**left column**) and ipsilateral trials (**right column**).



White dotted vertical line indicate lever presentation. Colorbars are provided for each row for each example session.

Figure 2.3.3: Mixed effect model regression on GCaMP6f traces of VTA/SN::DMS terminals (n = 12 recording sites) using Q values of contralateral minus ipsilateral. Same analysis as **Figure 2.3c**, except explanatory variables include the action of the mice (blue), the difference in Q values for contralateral and ipsilateral choices (orange), their interaction (green), and an intercept. Colored fringes represent +/- 1 standard error from estimates (n = 12 recording

sites). Dots at bottom mark timepoints where the corresponding effect is significantly different from zero at p<.05 (small dot), p<.01 (medium dot), p<.001(large dot). P values were corrected with Benjamini Hochberg procedure.



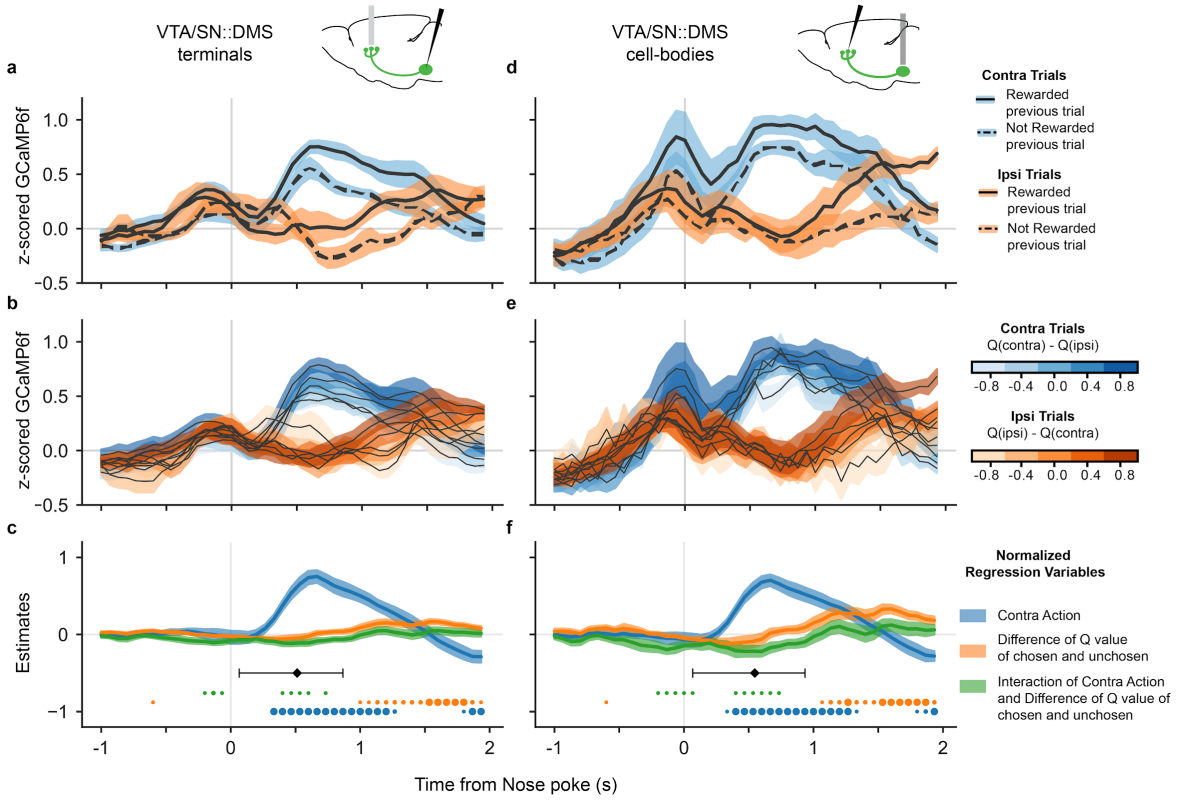


Figure 2.3.4: Analysis of DA signals time-locked to nose poke. (a) GCaMP6f signal time-locked to nose poke for contralateral trials (blue) and ipsilateral trials (orange), as well as rewarded (solid) and non-rewarded previous trial (dotted) from VTA/SN::DMS terminals. Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n =

12). **(b)** GCaMP6f signal for contralateral trials (blue) and ipsilateral trials (orange), and further binned by the difference of Q values of chosen and unchosen action. Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n = 12). **(c)** Mixed effect model regression on each datapoint from 3 seconds of GCaMP6f traces. Explanatory variables include the action of the mice (blue), the difference in Q values for chosen vs unchosen actions (orange), their interaction (green), and an intercept. Colored fringes represent +/- 1 standard error from estimates (n = 12 recording sites). Black diamond represents the average latency for lever presentation from nose poke, with the error bars showing the spread of 80% of the latency values. Dots at bottom mark timepoints when the corresponding effect is significantly different from zero at $p<.05$ (small dot), $p<.01$ (medium dot), $p<.001$ (large dot). P values were corrected with Benjamini Hochberg procedure. **(d-f)** Same as **(a-e),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

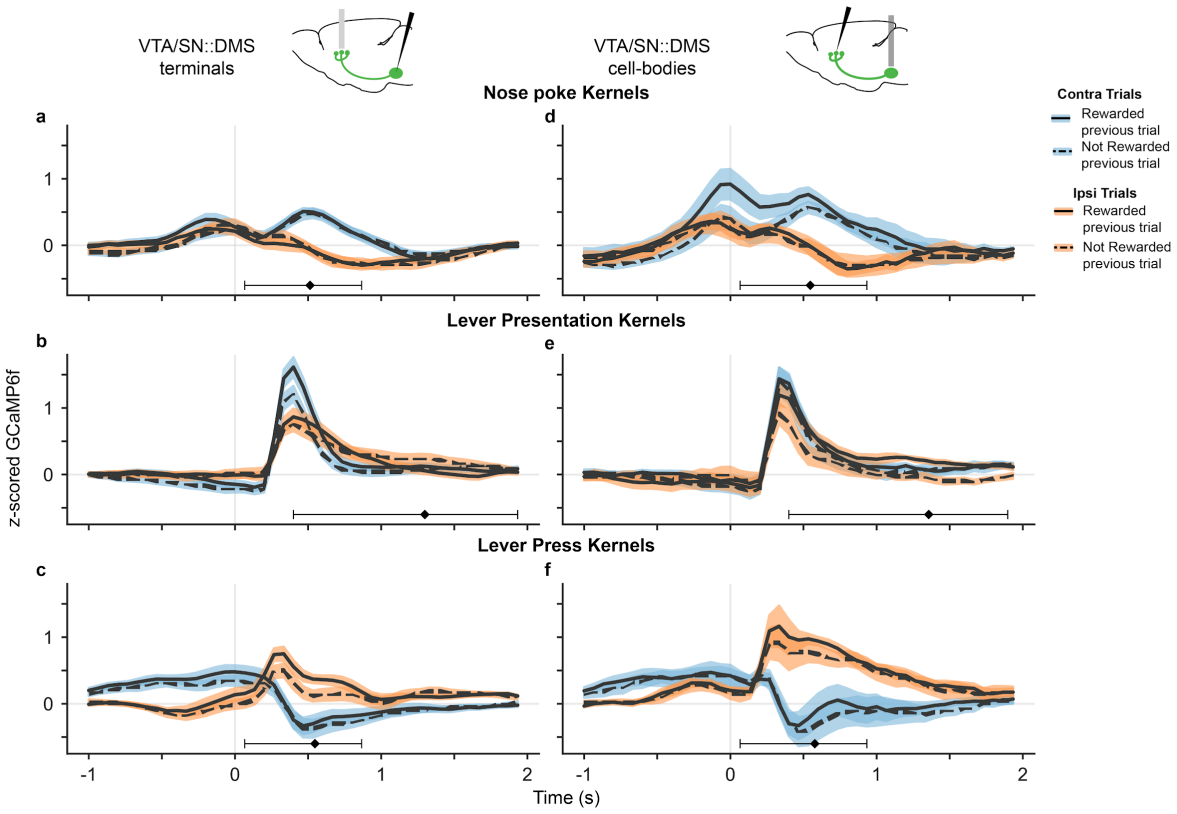


Figure 2.3.5: Kernels for each significant behavioral event from the multiple event kernel analysis (a) Nose poke kernel output from linear regression model using GCaMP6f from VTA/SN::DMS terminals. Each line is the kernel for a combination of contralateral (blue) and ipsilateral (orange) trials, as well as rewarded (solid) and non-rewarded (dotted) trials. Colored fringes represent +/- 1 standard error from activity averaged across recording sites (n = 12). Black diamond represents the average latency for lever presentation from nose poke with the error bars showing the spread of 80% of the latency values. **(b)** Lever presentation kernels, with the black diamond representing the average latency from lever press to lever presentation. **(c)** Lever press kernels, with the black diamond representing the average latency from CS+ or CS- to lever press. **(d-f)** Same as **(a-e),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

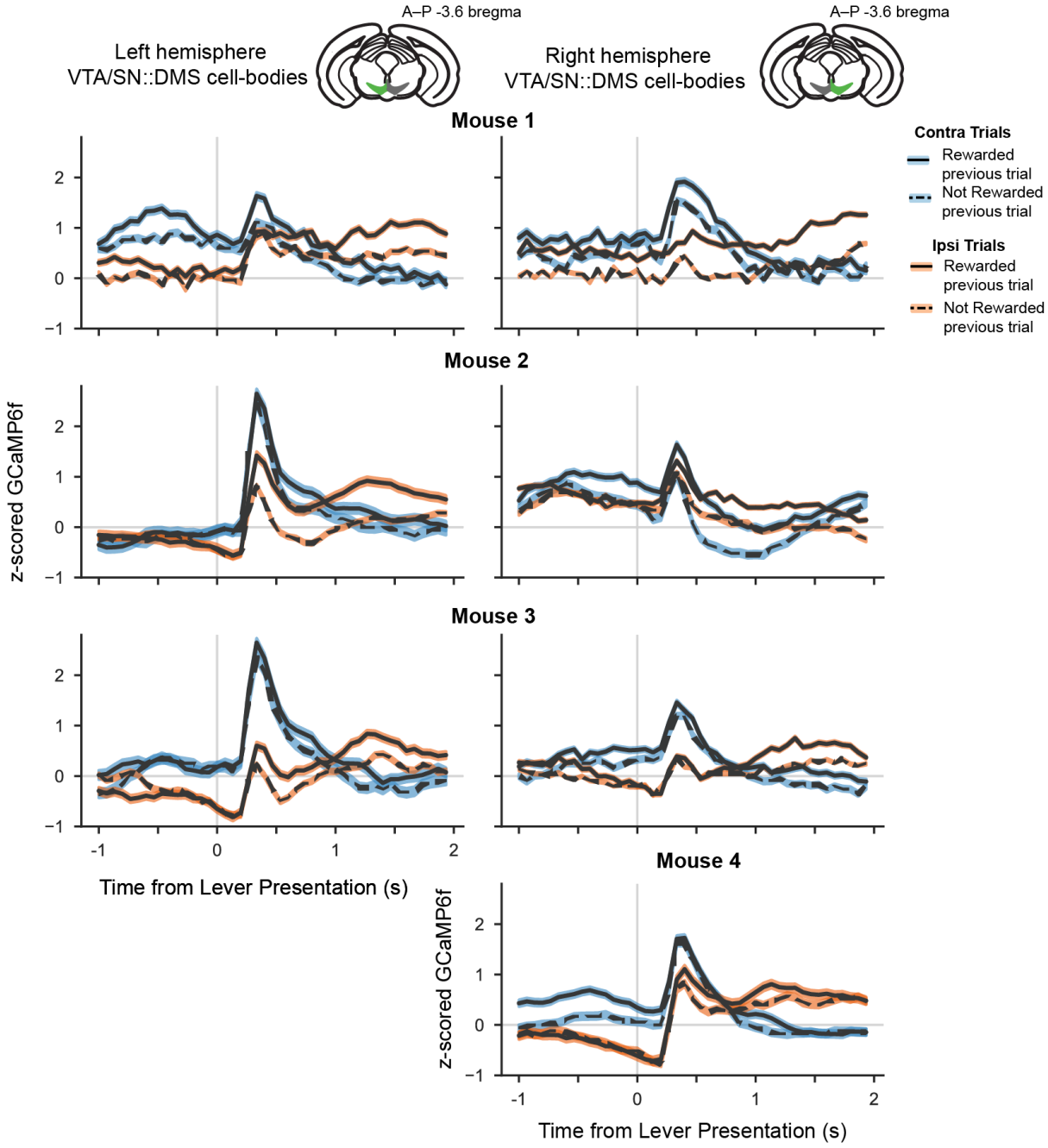

A–P -3.6 bregma
Left hemisphere
VTA/SN::DMS cell-bodies
A–P -3.6 bregma
Right hemisphere
VTA/SN::DMS cell-bodies
Mouse 1
Mouse 2
Mouse 3
Mouse 4
Contra Trials
Rewarded previous trial
Not Rewarded previous trial
Ipsi Trials
Rewarded previous trial
Not Rewarded previous trial
z-scored GCaMP6f
Time from Lever Presentation (s)

Figure 2.3.6: Averaged GCaMP6f signals of left and right hemispheres recordings from VTA/SN::DMS cell-bodies data (n = 4 mice, 7 recording sites). GCaMP6f signal relative to the lever presentation time for contralateral trials (blue) and ipsilateral trials (orange), as well as rewarded (solid) and non-rewarded previous trial (dotted) from VTA/SN::DMS terminals. Colored fringes represent +/- 1 standard error from activity averaged across trials. Each row represents averaged data from a distinct mouse (n = 4 total), with left and right column representing the left and right hemisphere recordings.

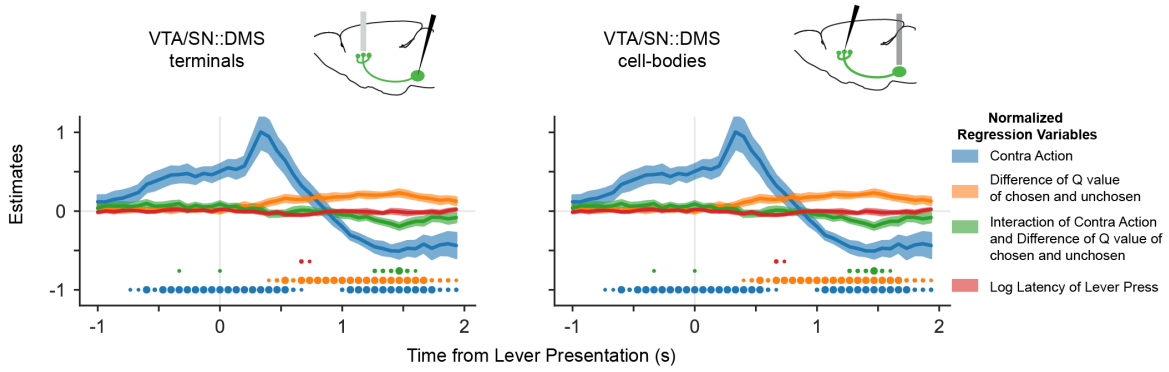


Figure 2.3.7: Mixed effect model regression with latency as nuisance covariate. (a) Mixed effect model regression with log latency of lever press (red) as additional nuisance covariate for VTA/SN::DMS terminal data (n = 12 recording sites). As with in **Figure 2.3c, f**, the mixed effect model's other explanatory variables include the action of the mice (blue), the difference in Q values for chosen vs unchosen actions (orange), their interaction (green), and an intercept. Colored fringes represent +/- 1 standard error from estimates. Dots at bottom mark timepoints when the corresponding effect is significantly different from zero at $p<.05$ (small dot), $p<.01$ (medium dot), $p<.001$ (large dot). P values were corrected with Benjamini Hochberg procedure. **(b)** Same as **(a),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

2.6 Acknowledgements

We thank the entire Witten and Daw labs for comments, advice and support on this work. I.B.W. is a New York Stem Cell Foundation—Robertson Investigator.

Chapter 3

A Feature-Specific Prediction Error Model Explains Dopaminergic Heterogeneity

The contents of this chapter have been published as:

Lee, Rachel S., Yotam Sagiv, Ben Engelhard, Ilana B. Witten, and Nathaniel D. Daw. 2023. "A Feature-Specific Prediction Error Model Explains Dopaminergic Heterogeneity." bioRxiv. https://doi.org/10.1101/2022.02.28.482379.

3.1 Introduction

Among the more prominent hypotheses in computational neuroscience is that phasic responses from midbrain dopamine (DA) neurons report a reward prediction error (RPE) for learning to predict rewards and choose actions (Schultz, Dayan, and Montague 1997; Houk, Adams, and Barto 1995; Montague, Dayan, and Sejnowski 1996). While clearly stylized, this account has impressive range, connecting neural substrates (the spiking of individual neurons and plasticity at target synapses, e.g. in striatum (Reynolds, Hyland, and Wickens 2001)) to behavior (trial-by-trial adjustments in choice tendencies (Parker et al. 2016)), all via interpretable computations over formally defined decision variables (**Figure 3.1a**). The question of this article is whether and how the strengths of this account can be reconciled with a growing body of evidence challenging a core feature of the model, i.e. the identification of a scalar, globally broadcast RPE signal in DA responses.

This scalar RPE is not a superficial claim of these theories, but instead one that connects a key computational idea to a number of empirical observations. Computationally, scalar decision variables reflect the ultimate role of any decision system in comparison: the decision-maker must order different outcomes against one another in 'common currency' to choose which to take (Von Neumann and Morgenstern 1947). Anatomically, the ascending DA projection has historically been argued to have an organization more consistent with a scalar-like "broadcast" code (rather than a labeled line code): a relatively small number of individual neurons innervate a large area of the forebrain via diffuse projections (Arbuthnott and Wickens 2007; Friston et al. 1994; Matsuda et al. 2009). Physiologically, early reports also stressed the homogeneity of responses of midbrain DA neurons on simple conditioning tasks – especially that a large majority of units

respond to unexpected reward (Schultz 1998).

However, the physiological argument for a scalar RPE is increasingly untenable, as a mounting body of recent work challenges the generality of this finding by demonstrating a range of variation in dopamine responses. Midbrain DA neurons can have heterogeneous and specialized responses to task variables during complex behavior (Parker et al. 2016; R. S. Lee et al. 2019; J. Y. Choi et al. 2020; Engelhard et al. 2019; Lerner et al. 2015; A. L. Collins and Saunders 2020; Verharen, Zhu, and Lammel 2020; Hassan and Benarroch 2015; Marinelli and McCutcheon 2014; Kremer et al. 2020; Howe and Dombeck 2016; Anderegg, Poulin, and Awatramani 2015; Barter et al. 2015; Cai et al. 2020; Hamid, Frank, and Moore 2021; Wei, Mohebi, and Berke 2021; Zolin et al. 2021), even while often having relatively homogenous responses to reward (Engelhard et al. 2019; Eshel et al. 2016; Schultz, Dayan, and Montague 1997; Cohen et al. 2012; Kremer et al. 2020). In some areas, the signature response to reward can itself be blunted or reversed (Howe and Dombeck 2016; Parker et al. 2016; Menegas et al. 2018; Jin and Costa 2010; Lerner et al. 2015).

How can we reconcile such DAergic heterogeneity with the substantial evidence for the RPE theory? There are a number of previous computational approaches relevant to this question, the vast majority of which share a common core hypothesis: that a set of different error signals could each serve to train different target functions, either predicting different aspects of reward (Daw, Kakade, and Dayan 2002; Lloyd and Dayan 2016; Dabney et al. 2020; Akiti et al. 2022), or predicting altogether different quantities (Rao 2010; Lak et al. 2017; Daw, Courville, and Touretzky 2003; Kurth-Nelson and Redish 2009; Gardner, Schoenbaum, and Gershman 2018). For example, distributional RL suggests prediction errors (PEs) to predict different quantiles of reward expectation (Dabney et al. 2020), action prediction error models envision PEs to predict specific

actions (Greenstreet et al. 2022; Bogacz 2020; Lindsey and Litwin-Kumar 2022), successor representation models suggest PEs for predicting different sensory stimuli (Gardner, Schoenbaum, and Gershman 2018), and threat prediction error models suggest a PE to predict threat (Akiti et al. 2022). We refer to this class of models collectively as "outcome-specific PEs," because they substitute a set of different prediction targets in place of the reward outcome in classic RPE (**Figure 3.1b**).

Two implications of this class of explanations limit its applicability. First, the substitution of different targets in place of reward means these models cannot explain why most DA neurons respond to reward, nor how such uniformity can coexist with heterogeneity in other aspects of responding (Engelhard et al. 2019). Second, anatomically, this approach amounts to duplicating the original scalar RPE model (**Figure 3.1a**) for multiple types of error signals in parallel circuits (**Figure 3.1b**). The fact that different targets need to be associated (at least roughly) in closed-loop fashion with their distinct PEs strongly suggests that DA neurons involved in different predictions must project to different target sites.

In contrast, here we propose a complementary new "feature-specific PE model" that offers an explanation for a distinct set of phenomena associated with heterogeneity *within* a projection-defined DA population, rather than across projections (**Figure 3.1c**). In our new model, each DA neuron calculates a PE based on the subset of inputs it receives, with the full PE signal reconstructed at the target site (**Figure 3.1c**). The key motivating observation is that within a single PE circuit (either the classic RPE, **Figure 3.1a**, or any of its siblings in the outcome-specific models, **Figure 3.1b**), the predictions being learned depend on an input variable known in RL models as "state": a high-dimensional group of all sensory and internal variables relevant to the

prediction. Since these are likely widely distributed throughout the brain, it is anatomically implausible that they converge uniformly and homogeneously on DA neurons to produce a scalar PE. Although the ascending DAergic projection is relatively diffuse (Haber, Fudge, and McFarland 2000), inputs to DA neurons are not homogenous but instead arise from cortico-basal-ganglionic circuits that are highly topographically organized (Alexander, DeLong, and Strick 1986; Lau, Monteiro, and Paton 2017; Haber, Fudge, and McFarland 2000; Hintiryan et al. 2016; Hunnicutt et al. 2016; Pan, Mao, and Dudman 2010; Cox and Witten 2019). Here we propose that a distributed code for state is carried by corticostriatal circuits, which transform it into corresponding distributed codes for value and RPE that, in effect, decompose these scalar variables over state features (Lau, Monteiro, and Paton 2017). In this way, different striatal and DAergic neurons reflect the contribution of different state features to value and RPE, but the ensemble collectively represents canonical RL computations over the scalar variables. In turn, this core circuit may be repeated in parallel between different targets, producing distinct aspects of heterogeneity between versus within projection-defined populations.

We initially investigate the feature-specific PE model by focusing on a recent study from our labs (Engelhard et al. 2019) which provides one of the most detailed and dramatic examples of variability at the level of single neurons within the VTA, the nucleus classically most identified with an RPE. By performing 2-photon imaging while mice performed an evidence accumulation task in a virtual reality T-maze, we observed that while neurons respond relatively homogeneously to reward during the outcome period, during the navigation and decision period, they respond heterogeneously to kinematics, position, cues, and more (Engelhard et al. 2019). We show how our feature-specific RPE model explains these data, while prominent outcome-specific models

cannot. Moreover, we demonstrate that our model can be integrated with the outcome-specific PE framework to explain aspects of SNc DA neuron cell-to-cell heterogeneity (Jin and Costa 2010).

Our new model offers a number of key insights. First, the contrast between the feature-specific and outcome-specific models establishes a broad framework predicting that distinct outcome- vs feature-related aspects of DAergic heterogeneity predominate between- vs within projection-defined DA populations with different targets, which immediately suggests direct experimental tests. Second, within the canonical VTA RPE circuit, the feature-specific model explains a striking contrast observed in dopamine neurons: heterogeneous responses to task variables alongside much more uniform outcome-period responses (Engelhard et al. 2019; Kremer et al. 2020). This is a key empirical signature of the distributed value code we posit, but poorly explained by the earlier class of explanations. Third, the model retains an algebraic mapping to the standard theory (Montague, Dayan, and Sejnowski 1996; Schultz, Dayan, and Montague 1997; Houk, Adams, and Barto 1995), preserving its successes while improving its match to anatomical and physiological evidence. Finally, the theory exposes an unexpected connection between the puzzling empirical phenomena of DAergic heterogeneity and a major theoretical question in RL models in neuroscience: the nature of state. While there has been substantial theoretical interest in the principles by which the brain represents task state (Wilson et al. 2014; Gershman, Blei, and Niv 2010; Y. K. Takahashi and Schoenbaum 2016; Y. K. Takahashi et al. 2011), there is relatively little empirical evidence to constrain these ideas. The new model suggests that the DA population representation itself can provide a window into this hitherto elusive concept of the neural representation of state.

3.2 Methods

Behavioral Task

Simulations: At every trial, the agent was placed at the start of a virtual T-maze, with cues randomly appearing on either side of the stem of the T-maze as the agent moved down the maze. On each trial, one side was randomly determined to be correct, and the number of cues on each side was then sampled from a truncated Poisson distribution, with a mean of 2.29 cues on the correct side and 0.69 cues on the incorrect side. In order to prevent the agent from forming a side bias, we used a debiasing algorithm to ensure that the identity of the high probability side changed if the agent kept choosing one side (Pinto et al. 2018). To match the procedure from the mouse experiment, we also oversampled easy trials (trials in which only one side had 6 total cues, and the other had no cues) by ensuring that they were 5% of the trials. The agent moved down the maze at a constant speed of 0.638 cm per timestep, and could also modulate its view angle with two discrete actions corresponding to left and right rotation. (A third discrete action moved forward without changing the view angle.) The cue region was 85 cm, and the cues were placed randomly along the cue region under a uniform distribution, but with the restriction that cues on either side had a minimal spatial distance of 14 cm between them. Each cue first appeared when the agent was 10 cm from the cue location and disappeared once the agent passed the cue by 4cm. After the cue region, there was a short, 5 cm delay region before the agent's final left or right action determined their choice of entering either arm in the T-maze. If the agent turned to the arm on the side where more cues appeared, they received a reward. The agent was also given a sensory input in the model indicating whether it made the correct or wrong turn.

Neural data: The task simulated above was a streamlined version of that used in the mouse recordings in Engelhard et al. (Engelhard et al. 2019) In particular, the rules for spacing and visual appearance of cues were the same, but the simulated controls were simplified (to discrete actions) and the maze was shorter (to facilitate neural network training). Thus the mice could control their speed and direction of movement more continuously by running on a trackball, and in this way traversed a maze that had a 30 cm start region (with no cues), a 220 cm cue region, and an 80cm delay region before the T maze arms. The mean numbers of cues were correspondingly larger: 6.4 on the correct side and 1.3 on the incorrect side. At reward time, the mice received a water reward if they made the correct choice; if a mouse made an incorrect choice, it was given a pulsing 6-12 kHz tone for 1 second. Before the next trial, the virtual reality screen froze for 1 second during reward delivery, and blacked out for 2 seconds if the mouse was rewarded or 5 seconds if the mouse failed.

Virtual Reality System and Deep Reinforcement Learning Model

A deep reinforcement learning network was trained on the evidence accumulation task. As input, the network took in 68 by 120 pixel video frames in grayscale. The model had 3 convolution layers to analyze the visual input, an LSTM layer to allow for memory, and output to 3 action units and 1 value unit. The first convolutional layer had 64 filters, a filter size of 8 pixels, and a stride of 2 pixels; the second convolutional layer had 32 filters, a filter size of 2 pixels, and a stride of 1 pixel; the third convolutional layer had 64 filters, a filter size of 3 pixels, and a stride of 2 pixels. The convolution layers fed into a fully connected layer, which fed into the LSTM layer with 64 units along with a second input, a one-hot vector of length 2 which flagged whether or not the agent was rewarded the end of the trial. The reward input into the LSTM was meant to replicate the

sensory input that the mouse experienced when it was rewarded with water or received a tone for failing the trial. The hyperparameters for the convolutional layers were optimized with a grid search of various filter numbers and sizes trained on supervised learning for recognizing towers. The convolution layer feeds into a fully connected feature layer of 128 units, then to 64 LSTM feature units, 64 value units, and 64 feature-specific RPE units.

We use the same MATLAB virtual reality program (ViRMEn software engine(Aronov and Tank 2014)) from the original neural recordings (Engelhard et al. 2019), which we altered to accommodate the agent's movement choices of forward, left, and right. While in the stem of the T-maze, the agent always moved forward at a constant rate per timestep. The constant speed ensured that at every trial, the agent always took the same number of timesteps to traverse the stem of the T-maze. The agent could choose to rotate left or right, which would alter the view angle 0.05 rads up to the limit of $-\pi/6$ and $\pi/6$ rads. The agent could also choose to move forward without changing their view angle. After the delay region in the T-maze, the agent's left and right movement would no longer alter its view angle, but instead determined which arm the agent chose.

In order for the deep RL agent to interact with the ViRMEn software, we created a custom gym environment using OpenAIGym's *gym* interface (Brockman et al. 2016). Our custom virtual reality gym environment defined the forward, left, and right movements the agent could make, and sent in the movement choices to the ViRMEn software which in turn returned updated video frames.

We trained the network to maximize obtained reward using the Stable Baselines (Hill et al. 2018) (version 2.10.1) implementation of the Advantage Actor Critic (A2C) algorithm (Mnih et al.

2016). The loss function for training using the A2C algorithm with weight parameters θ, action at time t a_t, state s_t and reward r_t is given as:

$$Loss(\theta) = log\, \pi(a_t \mid s_t, \theta)[A(s_t, a_t|\theta)] + \beta(V(s_t|\theta) - G(s_t|\theta)) - \eta H[\pi(\cdot\,|s_t, \theta)]$$

The first term is the actor loss, given as the log of the policy π multiplied by a sample of the advantage $A(s_t, a_t|\theta) = G(s_t|\theta) - V(s_t\,|\theta)$, where $G(s_t|\theta) = \sum_{i=0}^{n_{steps}-1} \blacksquare [\gamma^i r_{t+i}] + \gamma^{n_{steps}} V(s_{t+n_{steps}}|\theta)$ is an n - (defined as n_{steps}in the model) step Bellman estimator for the return and V is the value (critic) function. The second term is the critic loss, or the squared error of the value function estimate with respect to the return G. The final term is an entropy term which regularizes the policy to increase exploration. β and η are hyperparameters (listed below) to trade-off the effects of the regularization and the losses.

All hyperparameters used for the deep RL agent can be found in Table 1. We trained the model until it reached a performance of 80% or higher correct choices, which took 140.8 million timesteps or approximately 900,000 trials. To show that the results were consistent across multiple starting weights, we retrained the agent with fresh random weights and found that the new features replicated the main results of the paper (**Figure 3.9**).

We also ran a few additional simulations to learn more about the training process. First, pretraining the convolutional layers of the network to recognize towers significantly reduced training time, as well as shortening the delay region from the final tower appearance to the final tower location. The agent was also sensitive to the n_{steps} hyperparameter, which determined the batch gradient size; specifically, a shorter n_{steps} early in training helped with the agent learning

the task above chance, but longer n_{steps} was needed to master the task to reach the performance we see in the paper. We also trained the network with only 8 LSTM units, but the agent could not learn the task even after 200 million timesteps. A larger and overparameterized network was evidently needed for the agent to understand the task and keep track of the towers seen, even though the theoretical state space sufficient for the task is three dimensional (left tower count, right tower count, and position of the agent).

After training, we froze the weights and took the final output layer of the network before the action and value units (i.e., the LSTM output) as the features for vector value and feature-specific RPE (**Equation 3.1**, **Figure 3.2b**). (Note that this just corresponds to decomposing the scalar value and RPE units in the original A2C network into vectors, with one value component for every LSTM-to-value weight, and a corresponding RPE component: i.e. the feature-specific RPE model, being algebraically equivalent to TD, is just a more detailed view of the A2C critic). In this way, we calculated the feature-specific RPE at every point in the trial, for a total of 5000 simulated trials. We defined the outcome period to be the 5 timesteps before and after the reward. We defined the cue period as the first 140 timesteps of the maze, which occurred at the same positions on every trial since the agent always moved forward the same amount at each timestep.

Table 3.1: Hyperparameters for A2C Algorithm

Parameter (Variable Name in Stable Baselines)	
Number of Parallel Environments (n_env)	8

Number of steps until each environment updates (n_{steps})	140
Discount Factor (gammaγ)	0.99
Learning Rate	0.0025
Value function coefficient for loss calculation (vf_coef β)	0.25
Entropy coefficient for loss calculation (ent_coef η)	0.01
Maximum value for gradient clipping (max_grad_norm)	0.5
RMSProp decay parameter (alpha)	0.99
RMSProp momentum parameter (momentum)	0.0
RMSProp epsilon (epsilon)	1e-5

Neural Data

This article analyzes data originally reported in Engelhard et al (Engelhard et al. 2019), the methods of which we briefly summarize here and below (Engelhard et al. 2019). We primarily re-analyzed the neural recordings during the virtual-reality experiments, in which we used male

DAT::cre mice (n=14, Jackson Laboratory strain 006660) and male mice that are the cross of DAT::cre mice and GCaMP6f reporter line Ai148 (n=6 Ai148xDAT::cre Jackson Laboratory strain 030328).

VTA DA neurons were imaged at 30 Hz using a custom built, virtual-reality compatible two photon microscope equipped with pulsed Ti:sapphire laser (Chameleon Vision, Coherent) that was tuned to 920 nm. After imaging, we removed trials in which mice were not engaged in the task, primarily those found close to the end of the session when animal performance typically decreased. Average performance across sessions on all trials was 77.6 +/- 0.9% after removing trials (compared to 73.3 +/- 1.1% including all trials). Ultimately, we used 23 sessions from 20 mice (one session per imaging field, each session with at least 100 trials and minimal performance of 65%).

After preprocessing the imaging data, we performed motion correction procedures to eliminate spatially uniform motion and spatially non-uniform, slow drifts. The dF/F was derived by subtracting the scaled version of the annulus fluorescence from the raw trace (correction factor of 0.58) and smoothed using a zero-phased filter with 25 point center Gaussian with 1.5 sample points standard deviation. We then divided dF/F by the eighth percentile of the smoothed and neuropil-corrected trace based on the preceding 60s of recording. After examining the dF/F, we only included neurons that were stable for at least 50 trials. The full dataset we used for renalaysis has 303 neurons spread across 23 sessions from 20 mice.

Cue Period Responses

For the heatmaps in **Figure 3.2e-j**, each row represents a feature-specific RPE unit or neuron's response to the behavioral variable rescaled so that their maximum value is at 1 and minimum value is at 0. The example plots above show the unnormalized response. For the neural data, we used the same encoding model from our previous work (Engelhard et al. 2019) to predict neural activity with a linear regression based on predictors including cues, accuracy, previous reward, position, and kinematics, to isolate temporal kernels that reflect the response to each cue. The code for this encoding model can be found at https://github.com/benengx/encodingmodel. To determine which neurons are displayed for the heatmaps in **Figure 3.2h-j,** we used the same criterion as in Engelhard et al (Engelhard et al. 2019), which was to include neurons with a statistically significant contribution of that behavioral variable in the full encoding model relative to a reduced model, based on an F-test ($P = 0.01$), with comparison to null distributions produced by randomly shifted data to account for slow drift in the data.

Encoding Model and Clustering Analysis

For **Figure 3.11**, we adapted the encoding and clustering models from Engelhard et al (Engelhard et al. 2019) to estimate the relative contributions of various behavioral cues and variables to the model's simulated DA responses, and then cluster them using a Gaussian mixture model. First, we used multiple linear regression to model cue period activity, per simulated neuron, as a function of three behavioral features: cues, position, and the agent's choice. Each feature was coded with a group of regressors; we used the regression to measure the relative contribution (defined in terms of variance explained) of each group to each neuron's responses, for clustering (below). Coding of the variables followed that used for dopamine neurons in Engelhard et al (Engelhard et al. 2019): specifically, the cue variables were coded as event-locked timeseries, with the appearance of left

and right cues convolved with an 8 degrees of freedom regression spline basis across 10 timesteps of duration. Position was entered as a degree-3 polynomial series on the continuous variable, (i.e. we included the first, second, and third powers of position). We then chose the optimal number of predictors using a fivefold cross-validation over trials. The agent's final choice was coded with three variables as one-hot vectors, indicating the decision time for a left choice, and one and two steps before that. (There were no separate variables for choosing the right arm since we also included a bias variable in the regression.)

For the outcome period, we modeled timesteps following reward using two sets of regressors, representing (by one-hot event-triggered lagged indicators) the time of reward (on rewarded trials only) and time of outcome (comprising the time of either nonreward or reward). Each comprised 7 one-hot vectors indicating the event time (reward and/or nonreward) time and the 6 timesteps following. No bias variable was included.

After regressing cue and outcome period activity, the relative contribution and GMM clustering procedures were performed as in Engelhard et al (Engelhard et al. 2019). Following Engelhard, we define the relative contribution for each group of regressors by the reduction in variance explained when setting their betas to zero (referred to in Engelhard et al (Engelhard et al. 2019) as the "no refitting" method), then taking the ratio of these relative to the sum over all features.

Clustering was performed using MATLAB's `fitgmdist` function on the matrix of neuron x feature contributions, using 1,000 maximum iterations, 0.30 regularization value, 100 replicates, and the covariance matrix constrained to diagonal. To select the number of clusters, we used the fitgmdist function to calculate the AIC score for models with varying numbers of clusters and chose the

model with the lowest AIC score. The AIC score served as a measure of model fit given the number of parameters, calculated by taking the difference between 2 times the number of parameters and 2 times the log likelihood of the model. After taking the cluster with the lowest AIC scores, we ordered the relative contribution by those clusters.

Wall-Texture Analysis

In **Figure 3.3**, we identified a repeating wall-texture pattern in the maze by analyzing video frames of a maze with the view angle fixed at 0 degrees (N = 5000 trials). We calculated the similarity matrix for the video frames; specifically, given the video frames, we flattened the video frame at each timepoint into vectors, mean-corrected and normalized the vectors, and measured similarity for all pairs of frames as the cosine of the angle between these vectors (concretely, the i-jth entry of the similarity matrix gives the cosine of the angle between the video frames at times i and j). We also visualized the average, over positions, of each frame's similarity to those ahead and behind it, as a function of distance. We repeated the same analyses on the feature-specific RPEs, calculating the similarity in the feature-specific RPEs at each timepoint. The feature-specific RPEs here were derived by running the agent with the trained weights from the normal maze described above, but not allowing the agent to change its view angle in the stem of the maze (always fixed at 0 degrees). For **Figure 3.12**, we repeated the same analysis but calculated the similarity with the position lagged scalar RPE instead.

Confirmatory versus disconfirmatory cue responses

We defined confirmatory cues as cues that appeared on the side with more evidence so far, and disconfirmatory cues as cues that appeared on the side with less evidence so far. If the agent or

mouse had seen an equal number of cues on both sides, the next cue was defined as a neutral cue. For the neural data, we isolated cue kernels as in Engelhard et al (Engelhard et al. 2019) with some modifications: instead of using contralateral and ipsilateral cues, we used predictors including contralateral and ipsilateral cues with contralateral evidence, neutral evidence, and ipsilateral evidence so far. For **Figure 3.4b,** we selected those feature-specific RPE units that were modulated by cue onset and the 10 timesteps after cue onset, regardless of left or right cues. For **Figure 3.4c**, we selected among the neurons modulated by cues from the encoding model (N = 77/303), plotting only the units modulated by cues (N = 62/303).

For **Figure 3.13a-b,** we applied the same regression for **Figure 3.4c** to the original data split to recordings from the left and right hemispheres. For the feature-specific RPE model responses in **Figure 3.13c-d**, we split the units to left- versus right-cue responsive units based on the heatmaps in **Figure 3.2g**. The units were sorted by eye, choosing the units that responded to the left and right cues respectively at cue onset and throughout the 10 timesteps after cue onset and excluding units that responded at cue offset.

Outcome Period Responses

In **Figure 3.5a,d**, the scalar RPE was calculated by summing the feature-specific RPE units. For the model responses at outcome time for **Figure 3.5b,e**, we normalized each unit's activity 5 timesteps before and 5 timesteps after reward across 5000 trials so the minimum is 0 and the maximum is 1. Then, we averaged across trials and took the response at reward time. For the neural responses at outcome time for **Figure 3.5c,f**, we matched the original empirical paper (Engelhard et al. 2019) and calculated the average activity in the first 2 seconds after the onset

of the outcome period, baseline corrected by subtracting the average activity from the 1 second period preceding the outcome. For the histograms in **Figure 3.5b-c**, **e-f,** a two-sided Wilcoxon signed rank test was performed to determine the p value for the median (yellow line).

Successor Representation Modeling

We modeled a family of DA responses using a vector of prediction errors for a successor representation, based on Gardner et al (Gardner, Schoenbaum, and Gershman 2018). For comparability with the feature-specific RPE model, we used as targets for these successor predictions the same network-derived features $\vec{\phi}$ from the feature specific RPE model (LSTM units from **Figure 3.2b**). Thus, we have a successor feature model, whose weights we learned using Algorithm 3 of Barreto et al (Barreto et al. 2016). Given the input state $\vec{\phi}(s_t)$, we trained a set of weights W to learn a vector of successor features $\overrightarrow{SF}$ with each SF_i trained with its own sensory prediction error δ_i (**Figure 3.6a**). We z-scored the target features $\vec{\phi}(s_t)$ before training the successor feature model. To calculate the sensory prediction errors (to match with the feature-specific PEs in **Figure 3.1c**), we trained a weight matrix W with rows of $\vec{w}_i$, such that each successor feature

$$SF_{i,t} = \vec{\phi}_t \cdot \vec{w}_i \text{ [3.2].}$$

Each $\vec{w}_i$ for each successor feature SF_i was then updated using the classic scalar RPE recursive step; that is, each has its own sensory prediction error δ_i

$$\delta_{i,t} = \phi_{i,t} + \gamma SF_{i,t+1} - SF_{i,t} \text{ [3.3]}$$

We updated each $\vec{w}_i$ accordingly using:

$$\vec{w}_i \leftarrow \vec{w}_i + \alpha\delta_{i,t}\vec{\phi}_t [3.4].$$

A total of 64 successor features were trained to match with the vector of 64 input features $\vec{\phi}_t$. We trained using this recursive update for 15000 trials until the weights converged, using discount factor $\gamma = 0.99$ and a learning rate $\alpha = 0.001$.

After training our weight matrix W, we reran the algorithm with the learned weights frozen to calculate the 64 sensory prediction errors for the 64 features. In **Figure 3.6b-d,** we generated the heatmaps with the same methods as with feature-specific RPE model's results in **Figure 3.2e-g**. In our analyses in **Figure 3.6e,** we compare the feature-specific RPE model's scalar RPE with the averaged sensory PE, which is the average over the 64 sensory PEs**.** The outcome period plots **Figure 3.6f-i** was generated the same way as **Figure 3.5a-b,d-e**.

Distributional Reinforcement Learning Modeling

To model the features with a distributional reinforcement learning model, we adapted a TD-1 algorithm to train weights for 31 distribution channels, based on algorithms specified in refs. (Dabney et al. 2020; Rowland et al. 2019) and briefly summarized below. Each channel has two learning rates, ${\alpha_i}^+$ for positive prediction errors and ${\alpha_i}^-$for negative prediction errors, chosen so that the expectiles $\tau_i = \frac{{\alpha_i}^+}{{\alpha_i}^+ + {\alpha_i}^-}$evenly tile the range between 0 and 1 (exclusive) with ${\alpha_i}^+ + {\alpha_i}^- = 0.001$.

The input features for the model were (again, for comparability with the feature-specific RPE) that model's network-derived features $\vec{\phi}$, z-scored. We trained weights that mapped the features to distributional value channels, including a bias term. We trained a set of linear feature weights for each channel by at each timestep directly predicting the (γ= 0.99 discounted) final outcome

of the trial (effectively, expectile regression or TD-1; this avoids distributional imputation in a bootstrapped TD-0 update).

After training for 15000 trials, we froze the converged weights to calculate the 31 expectile prediction errors for the distribution channels. We then computed TD-0 Bellman errors at each step with an imputation algorithm (Rowland et al. 2019) to sample the one-step bootstrap backup value for each of the 31 channels from the ensemble.

To generate plots for **Figure 3.7b-i**, we repeated the same methods as in **Figure 3.6b-i,** with one difference for the heatmaps in **Figure 3.7b-d**; specifically, in the heatmaps, we used the original ordering of the distribution channels, since the channels themselves already have a natural order from pessimistic to optimistic channels.

Action Prediction Error Models

The Action Prediction Error (APE) framework may be formally thought of as the standard RL setup but execution of actions provides the reward. The scalar APE model used to describe the Parker et al (Parker et al. 2016) results correspondingly used a variant of the typical Q-learning approach. For a hypothetical DA population with some preferred action a (e.g., left lever press), that is:

$$\delta_t{}^a = I_{a_t=a} + \gamma V^a(s_{t+1}) - V^a(s_t) \quad [3.5]$$

$$V^a(s_t) \leftarrow V^a(s_t) + \eta\delta_t{}^a \quad [3.6]$$

Here, ${\delta_t}^a$ is the prediction error at time t and corresponds to the dopaminergic response. $I_{a_t=a}$ is an indicator for the population's preferred action: equal to 1 if $a_t = a$ and 0 otherwise. $\gamma \in [0, 1]$ is a temporal discounting factor. $V^a(s)$ denotes the value of state s, which is equal to the expected cumulative discounted number of times a will be executed during the remainder of the trial. η is a learning rate parameter.

The feature-based APE model used to describe Jin and Costa (Jin and Costa 2010) is similar. Indeed, the approach used to derive it was identical; we treated the actions as being the source of the rewards:

$$\delta^a{}_{i,t} = \frac{I_{a_t=a}}{N} + \gamma V^a{}_{i,t+1} - V^a{}_{i,t} = I_{a_t=a} + w^a{}_{i,t}(\gamma \phi^a{}_{i,t+1} - \phi^a{}_{i,t}) \quad [3.7]$$

$$w^a{}_t \leftarrow w^a{}_t + \eta \delta^a{}_t \phi^a{}_t \quad [3.8]$$

Here, $\delta^a{}_{i,t}$ is the prediction error at time t for feature channel i and corresponds to the dopaminergic response of that DA neuron. N is the total number of channels. $w^a{}_{i,t}$ is the ith entry in the value weight vector at time t. $\phi^a{}_{i,t}$ is the ith entry in the state vector at time t.

Our models of the Parker et al (Parker et al. 2016) task and the Jin and Costa (Jin and Costa 2010) task were essentially the same with some slight differences in parameters, so we will first describe their shared abstract structure and then elaborate separately on the parameters. Both tasks began at a "start" state in which levers were presented. In order to randomize the timing of subsequent actions, the start state had a probabilistic self-transition. After exiting the start state, the agent needed to press a lever a fixed number times in order to progress to the outcome states;

they chose which lever to press at each timepoint according to a softmax policy over the true reward values for each lever. For each press, they first progressed to a "premotor" state that preceded the chosen lever press state (i.e., led to it with probability 1 on the next timestep) and then to the "press" state itself (where the agent that preferred the corresponding action earned a reward of 1). The final choice state similarly had a probabilistic self-transition, after which the agent entered a reward state or an omission state depending on the levers it had pulled. An abstract state diagram may be found in **Figure 3.15**.

In the Parker et al (Parker et al. 2016) task, the agent only needs to press a lever once to reach the outcome states. The softmax policy has an inverse temperature T = 0.25. The start state self-transition has probability 0.8. The pre-outcome self-transition has probability 0.95. The temporal discount factor for the reward model is $\gamma_{RPE} = 0.95$. The temporal discount factor for the APE model is $\gamma_{APE} = 0.5$. Discounting is harsher for the APE model to match the value used in modeling the Jin and Costa (Jin and Costa 2010) task (described below) – but we note that the specific values used do not change the qualitative pattern of the results but only affect the scale of the responses in **Figure 3.8d, left**.

In our simplification of the Jin and Costa (Jin and Costa 2010) task, the agent needs to press a lever three times to reach the outcome states. To randomize the within-sequence press timing, each of the choice states also has a self-transition with probability 0.1. The start state self-transition has probability 0.8. The softmax policy has an inverse temperature $\tau = 0.25$. The temporal discount factor for both the scalar and feature-specific models is $\gamma = 0.5$. In general, we

believe that harsher discounting is appropriate for the APE models since the computational goal of action prediction is presumably preparing a single upcoming response rather than computing a long-run average– but we note, as above, that the specific value of γ does not affect the qualitative pattern of the results (i.e., that there are DA neurons with specific preferred presses that emit a biphasic response, and that those neurons will change their tuning in response to changes in tuning upstream), only the relative magnitudes of the peaks and troughs in the DA response.

Retraining Network Results

The retrained network results in **Figure 3.9** used the same hyperparameters as the original run and took 162 million timesteps to learn. One minor difference was the scheduling of the hyperparameter n_{steps}, which we lowered to 20 early in training, then increased to 50 for the last 28 million timesteps. The lower n_{steps} helped improve performance in part because it meant the A2C algorithm had fewer trajectories it needed to sample and explore and therefore could more easily learn the task. After performance was above chance, a higher n_{steps} was needed to match the original performance in order to allow the agent to integrate the cues across the entire trial to maintain the correct tower difference count. The plots for **Figure 3.9** were generated the same as their counterparts in **Figure 3.2c-d, Figure 3.4b, Figure 3.2e-g,** and **Figure 3.5a-b,d-e.**

3.3 Results

The Feature-Specific PE model

Here, we propose the feature-specific PE model as an extension of the classic scalar RPE model (**Figure 3.1a**). RL models typically assume that the goal of the learner is to learn the value function $V(s_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid s_t]$ (i.e. the expected sum of γ-discounted future

rewards r_t starting in some state s_t). Both in neuroscience and in AI (Sutton and Barto 2018), a typical starting assumption for high-dimensional or continuous tasks is to assume the learner approximates value linearly in some feature basis. That is, it represents the state s_t by a vector of features $\vec{\phi}(s_t)$ (henceforth, $\vec{\phi}_t$) and approximates value as a weighted sum of those features, i.e. $V_t = V(s_t) \approx \vec{\phi}_t \cdot \vec{w}$. This reduces the problem of value learning (for some feature set) to learning the error-minimizing weights $\vec{w}$ and, more importantly for us, formalizes the state representation itself as a vector of time-varying features $\vec{\phi}_t$. (The linearity assumption is less restrictive than it seems, since the features may be arbitrarily complex and nonlinear in their inputs. For instance, this scheme is standard in AI, including at the final layer in deep-RL models, which first derive features from a video input by multilayer convolutional networks, then finally estimate value linearly from them (Mnih et al. 2015, 2016).)

In a standard temporal-difference (TD) learning model, weights are learned by a delta rule using the RPE $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ (Sutton 1988). A typical cartoon of how these are mapped onto brain circuitry is shown in **Figure 3.1a**, with a cortical input population vector for state features projected to scalar value and RPE stages, corresponding, respectively, (since both variables are scalar) to presumed uniform populations of striatal and midbrain DA neurons (Houk, Adams, and Barto 1995; Montague, Dayan, and Sejnowski 1996; Schultz, Dayan, and Montague 1997). The RPE then drives weight learning at the corticostriatal synapses via ascending DA projections. Previously proposed models to account for DA heterogeneity often replace the target r_t with one or more additional (potentially non-reward) target outcomes, replicating this scalar prediction model across parallel circuits ("outcome-specific PE models"; **Figure 3.1b**)(Daw, Kakade, and Dayan 2002; Lloyd and Dayan 2016; Gardner, Schoenbaum, and Gershman 2018;

Dabney et al. 2020; Akiti et al. 2022; Rao 2010; Lak et al. 2017; Daw, Courville, and Touretzky 2003; Kurth-Nelson and Redish 2009; Gershman, Pesaran, and Daw 2009). These models can explain some types of variation in DA across projection-defined populations.

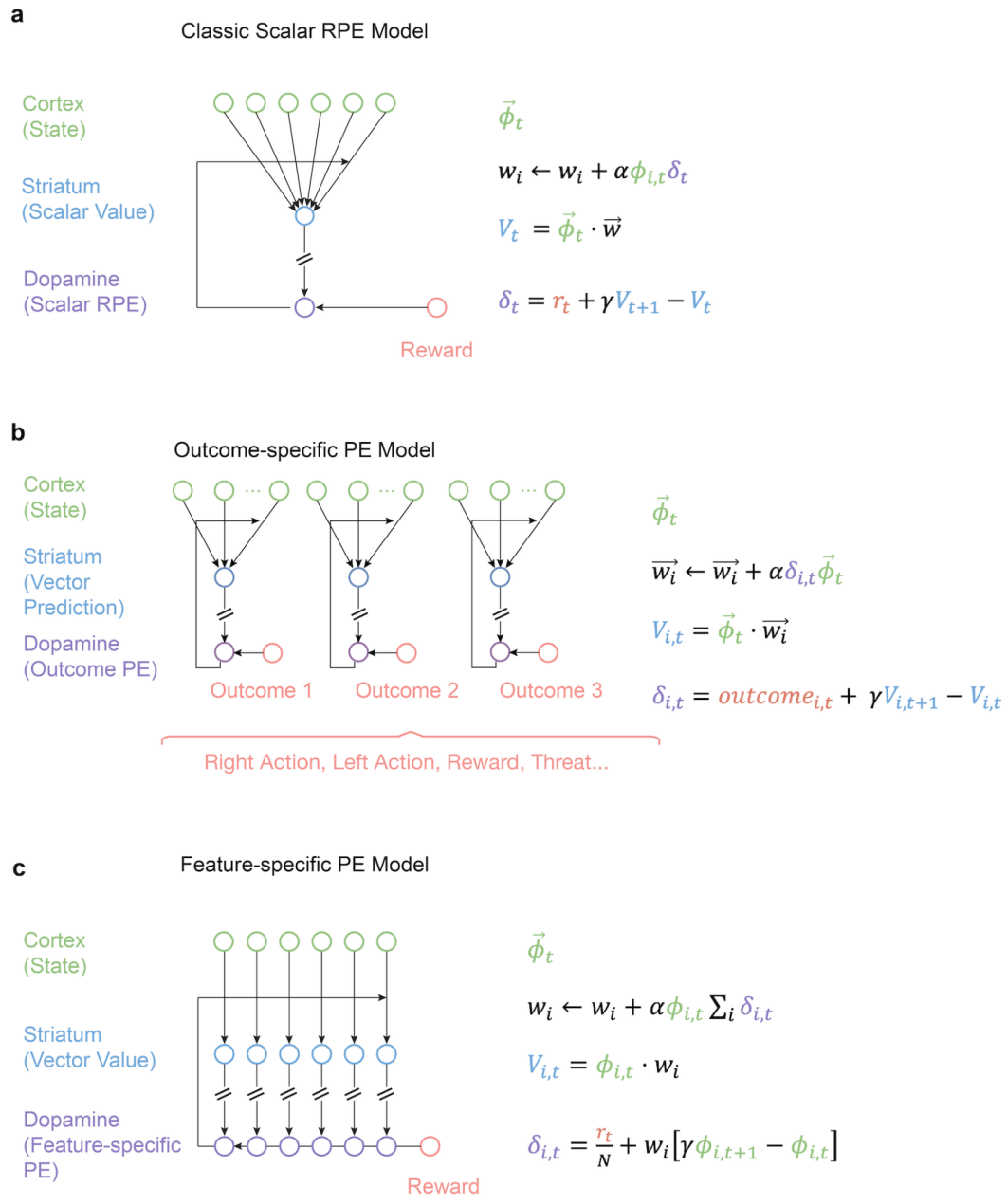


Figure 3.1: Feature-specific PE model updates classic TD learning to produce heterogeneous DA signals that reflect the state representation. **(a)** Classic mapping between equations of the

TD learning model and brain circuitry (Houk, Adams, and Barto 1995; Montague, Dayan, and Sejnowski 1996; Schultz, Dayan, and Montague 1997). **(b)** RL models can use different outcome variables to generate vector values and PEs. These outcomes can be driven by different aspects of reward (Daw, Kakade, and Dayan 2002; Akiti et al. 2022; Dabney et al. 2020) (like in distributional RL models), specific actions (such as nosepoke, left, and right action) (Greenstreet et al. 2022; Bogacz 2020; Lindsey and Litwin-Kumar 2022), or other targets (Rao 2010; Lak et al. 2017; Kurth-Nelson and Redish 2009; Daw, Courville, and Touretzky 2003; Gardner, Schoenbaum, and Gershman 2018). **(c)** Our proposed feature-specific PE model, which remaps the scalar RPE algorithm onto brain circuitry such that value and PE are vectors but the overall computations are preserved.

In contrast, we propose a new "feature-specific RPE model", which produces heterogeneity even within a projection-defined DA population. This model relaxes the unrealistic anatomical assumption that the corticostriatal stage of this circuit involves complete, uniform convergence from vector state to scalar value (**Figure 3.1c**). In fact, projections are substantially topographic at each level of the corticostriatal circuit(Alexander, DeLong, and Strick 1986; Haber, Fudge, and McFarland 2000; Voorn et al. 2004). Thus, if different striatal units i preferentially receive input from particular cortical features $\phi_{i,t}$, then value will itself be represented by a distributed feature code $V_{i,t} = w_i\phi_{i,t}$ (see also (Montague, Dayan, and Sejnowski 1996; Suri and Schultz 1999)), and, in turn, DA neurons preferentially driven by each "channel" will compute feature-specific reward prediction errors

$$\delta_{i,t} = \frac{r_t}{N} + \gamma V_{i,t+1} - V_{i,t} = \frac{r_t}{N} + w_i(\gamma\phi_{i,t+1} - \phi_{i,t}) \text{ [3.1]}$$

(where N is a scale factor equaling the number of channels). Importantly, due to linearity, the aggregate response (summed over channels i) at the value stage reflects the original scalar value, and the aggregate response at the RPE stage corresponds to the original scalar RPE.

Thus (assuming the ascending dopaminergic projection is sufficiently diffuse as to mix the channels prior to the weight update), the model corresponds algebraically to the classic one, just mapped onto the brain circuitry in a more realistic way. Our primary insight is that nonuniform convergence from the state layer to the value layer, and from the value layer to the dopamine layer, will result in value and RPE stages that reflect input feature variation, but preserve the scalar RPE model (**Figure 3.1a**) due to linearity. Note that this insight holds, at least approximately, under more realistic models in which channels partially mix at each step due to anatomically limited convergence (so that, for instance, a feature at the value or RPE stage is the average of several features from the afferent population) or in which linearity is only approximate.

While still quite stylized, this account represents a more realistic picture compared to the traditional scalar story of both value and RPE. Notably, it better reflects the known topography in the projections from cortical inputs to MSNs and from MSNs to dopamine units (Alexander, DeLong, and Strick 1986; Cox and Witten 2019; Lau, Monteiro, and Paton 2017), and physiological findings that both input populations show variable tuning to similar features as do the DA neurons (Rueda-Orozco and Robbe 2015; Tsujimoto, Genovesio, and Wise 2011; Parker et al. 2022; N. Matsumoto et al. 2001; Robbe 2018; K. Choi et al. 2023; Jin and Costa 2010). Incorporated in the feature-specific RPE model, (even without specifying a particular feature basis) these properties directly imply several aspects of the heterogeneous dopamine response. Consider times when primary reward is not present, such as during the navigation and decision-

making period in Engelhard et al (Engelhard et al. 2019) Here, $r_t = 0$ and Equation 1 reduces to $\delta_{i,t} = w_i(\gamma\phi_{i,t+1} - \phi_{i,t})$: that is, each DA unit reports the time-differenced activity in its feature, weighted by its own association w_i with value. Depending on what the features $\vec{\phi}_t$ are, this would explain dopaminergic response correlations with different, arbitrary covariates: to the extent some feature ϕ_i is task-relevant, it will have nonzero w_i (where the sign of w_i is determined by ϕ_i's partial correlation with value in the presence of the other features), and its derivative will correlate with a subset of neurons. Even a DA neuron that is driven by an objectively task-irrelevant feature is likely to respond to it, due to incidental or transient correlations between the feature and value producing nonzero w_i.

Conversely, at outcome time, all modeled RPE units are likely to respond differentially for reward than nonreward (due to r_t being shared across all channels in Equation 1) as in Engelhard et al (Engelhard et al. 2019). This reward response may also be modulated by its predictability, due to each channel's share of the temporal difference component, $w_i(\gamma\phi_{i,t+1} - \phi_{i,t})$, but is unlikely to be completely predicted away by most individual features. Finally, since the standard RPE δ_t is equal to the sum over all channels $\Sigma_i\delta_{i,t}$by construction, the model explains why neuron-averaged data (or bulk signals as in fiber photometry or BOLD), as often reported, resemble TD model predictions even potentially in the presence of much inter-neuron variation.

Deep RL network to simulate the Feature-Specific PE model

Although our feature-specific PE model is fully general, in order to simulate the model in the context of a specific task, we need to specify an appropriate set of basis functions for the task state. We consider our previously reported experiment in which mice performed an evidence

accumulation task in a virtual reality environment while VTA DA neurons were imaged (Engelhard et al. 2019) (**Figure 3.2a**). In this task, mice navigated in a virtual T-maze while viewing towers that appeared transiently to the left and right, and were rewarded for turning to the side where there had been more towers.

To simulate the vector of feature-specific PEs (in this task, RPEs), we took advantage of the fact that this task was based in virtual reality, and therefore we could train a deep RL agent on the same task that the mice performed to derive a vector of features, and in turn, a corresponding distributed code for values and RPEs (**Figure 3.2b**). In particular, we used a deep neural network to map the visual images from the virtual reality task to 64 feature units (via three convolutional layers for vision, then a layer of LSTM recurrent units for evidence accumulation; see Methods). These features were then used as common input for an actor-critic RL agent: a linear value-predicting critic (as above, producing the feature-specific RPE that sums to the traditional scalar RPE) and a softmax advantage learner responsible for choosing an action (left, right, forward) at each step. We trained the network to perform the task using the A2C algorithm (Mnih et al. 2016). Note that although A2C does not use the feature-specific RPEs individually for training, only their sum, this is also true in our model: the decomposition over features, in both cases, arises at an intermediate stage of computation.

After training, the agent accumulated evidence along the central stem of the maze to ultimately choose the correct side with accuracy similar to mice (shown as a psychometric curve in **Figure 3.2c; Figure 3.9** shows that this and later results are consistent in an independent training run). A minimal abstract state space underlying the task is 2D, consisting of the position along the maze and the number of towers seen (on left minus right) so far. When examining average

responses of the trained state features from the network in this space, we find that units are tuned to different combinations of these features, implying that they collectively span a relevant state representation for the task (**Figure 3.10**). Further, the scalar value function output by the trained agent while traversing the maze, derived by summing the value vector, is modulated by trial difficulty (operationalized, following the mouse study, as the absolute value of the difference in the number of cues presented on either side), meaning that the trained agent can predict the likelihood of reward on each trial (**Figure 3.2d**).

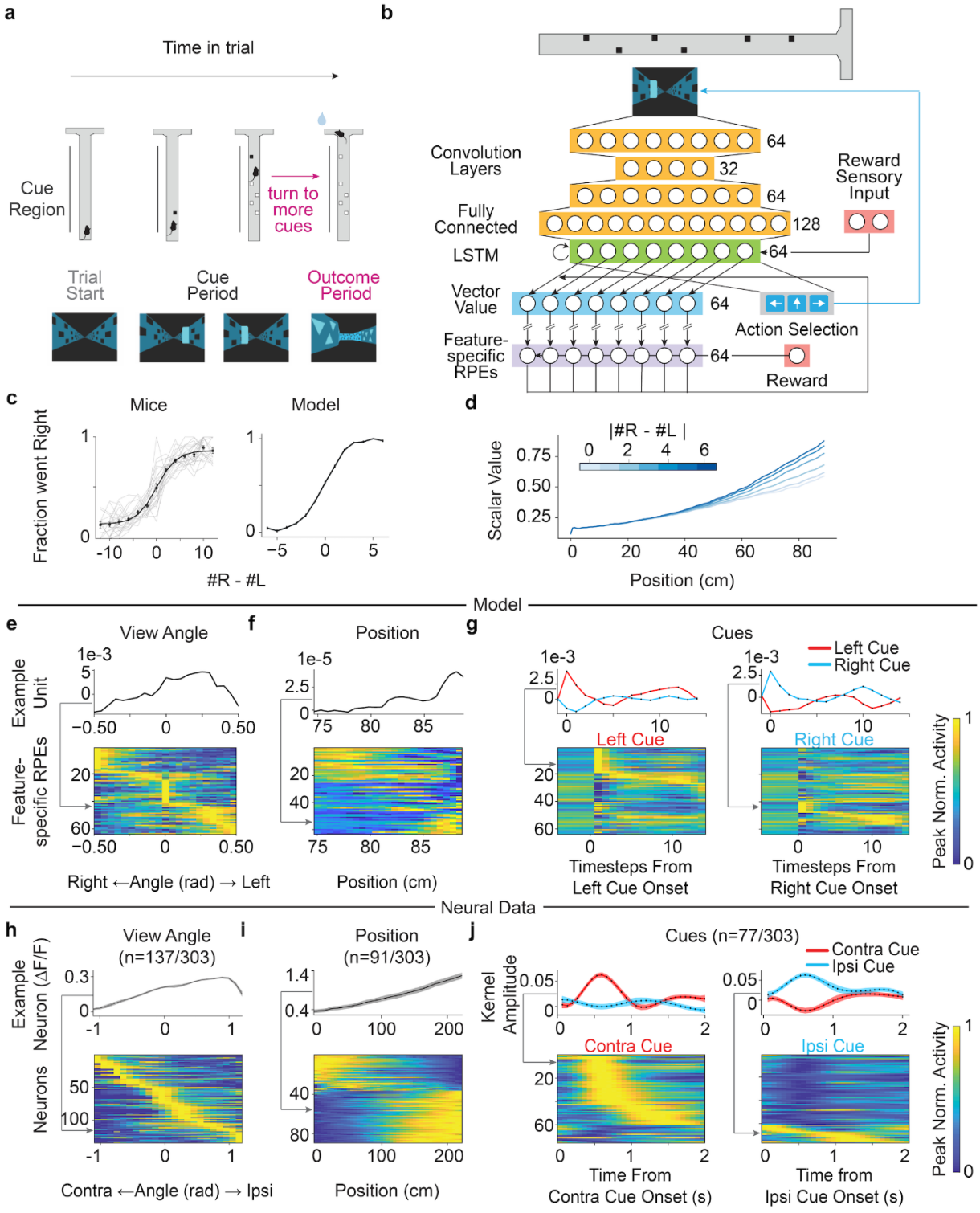


Figure 3.2: Feature-specific RPEs derived from a deep reinforcement learning model trained on a VR evidence accumulation task are heterogeneously modulated by behavioral variables

during the navigation and decision period, similar to DA neurons. **(a)** Task schematic of the VR task in which mice accumulated visual evidence (cues) as they ran down the stem of a T-maze and were rewarded if they turned to the side with more cues at the end. Video frames of the maze are shown below each maze schematic. **(b)** A deep reinforcement learning (RL) network took in video frames from the VR task, processed them with 3 convolution layers, a fully connected layer (orange), and an LSTM layer (green), and outputted an action policy (gray), which inputted the chosen action back into the VR system (blue arrow). The second to last layer of the deep RL network (the LSTM layer) and the weights for the critic served as the inputs to form the feature-specific RPEs (purple). Reward information (red) was inputted both as a one-hot sensory signal to the LSTM layers (for rewarded and unrewarded trials) and also to the feature-specific RPEs. **(c)** Psychometric curves showing the mice's performance (left) and agent's performance (right) after training. The fraction of right choices is plotted as a function of the difference of the right and left towers presented on the trial. For the mice, gray lines denote the average psychometric curves for individual sessions and the black line denotes a logistic fit to the grand mean with bars denoting the s.e.m. (N = 23 sessions). For the model, black bars indicate the s.e.m. (N = 5000 trials) **(d)** Deep RL model's scalar value (sum over units in vector value) during the cue period decreased as trial difficulty (measured by absolute value tower difference, blue gradient) increased. **(e)** Feature-specific RPE unit activity with respect to the view angle of the agent (min-max normalization; top: example unit; bottom: all units). **(f)** Same as **(e)** but for position of the agent in the final 15 cm of the maze. **(g)** Same as (**e-f**) but for left (red) and right (blue) cues. **(h-j)** Same as **(e-g)** but for the subset of neurons from Engelhard et al(Engelhard et al. 2019) tuned to **(h)** view angle of the mice, **(i)** position, and **(j)** contralateral (red) and ipsilateral (blue) cues. Fringes represent ±1 s.e.m. In **(h-i)** peak normalized $\Delta F/F$ signals are plotted while in **(j)** cue

kernels are plotted from an encoding model used to quantify the relationship between the behavioral variables and each neuron.

Feature-specific RPEs have heterogeneous selectivity during the cue period

The key finding from our prior cellular resolution imaging of DA neurons during the virtual T-maze task was heterogenous coding of task and behavioral variables during the navigation and decision period, such as view angle, position, and cue side (contralateral versus ipsilateral) (Engelhard et al. 2019). This heterogeneity was followed by relatively homogeneous responses to reward during the outcome period.

We first sought to determine if the feature-specific RPEs from our agent had heterogenous tuning to various behavioral and task variables during the cue period, similar to our neural data. We first considered the view angle during the central stem of the maze, which had no effect on reward delivery in the simulated task, but which the agent could nonetheless rotate by choosing left or right actions while in the stem of the maze. The feature-specific RPEs displayed idiosyncratic selectivity across units for the range of possible view angles (**Figure 3.2e**), which qualitatively resembled our previous results from DA neuron recordings (**Figure 3.2h**). A subset of RPE units showed position selectivity, including both downward and upward ramps towards the end of the maze (**Figure 3.2f**), again qualitatively resembling our neural recordings (**Figure 3.2i**). Finally, as in the neural data (**Figure 3.2j**), our feature-specific units also had idiosyncratic and heterogeneous cue selectivity, including preference for right vs left cues and diversity in the timing of the response (**Figure 3.2g**). Similar to the neural data (Engelhard et al. 2019), we observe clustering of these

feature responses (**Figure 3.11a**), with different clusters responding most strongly to cues, position, and choice, respectively (**Figure 3.11b**).

Reward-irrelevant responses in the feature-specific RPEs

A property of the neural data is the encoding of task features that appear to be reward-irrelevant (Engelhard et al. 2019). In the model, the requirement of the network to extract relevant task state from the high-dimensional video input implies the possibility that reward-irrelevant aspects of the input may "leak" into the state features, and ultimately into the feature-specific RPEs – even if they average out in the scalar RPE. Although we of course do not intend backpropagation as a mechanistic account of how the brain learns features, its use here exemplifies the more general principle that a low-dimensional output objective (choosing actions and predicting scalar reward) imposes few constraints on higher-dimensional upstream feature representations.

To test the validity of this idea, we sought to examine whether there may be coding of reward-irrelevant visual information in the feature-specific RPEs of the agent. We focused on unambiguously reward-irrelevant visual structure in the task, namely incidental background patterns that appear on the wall in the stem of the maze (**Figure 3.3a**). This background pattern repeats every 43 cm, a structure which is clearly visible as off-diagonal banding in the matrix of similarity between pairs of video frames across all combinations of locations. The structure is also visible as peaks in the 1D function (similar to an autocorrelation) showing the average similarity as a function of the distance between frames (**Figure 3.3b**). To investigate whether the same irrelevant feature dimensions are present at the level of the feature-specific RPEs, we repeat the same analysis on them. (To ensure the network inputs actually reflect such repeating similarity

structure, for this analysis we exposed the trained network to a maze traversal with a fixed view angle). The resulting feature-specific RPEs show the same pattern of enhanced similarity at the characteristic 43 cm lag, supporting the expectation that task-irrelevant features do propagate through the network (**Figure 3.3c-d**). This particular effect remains a prediction for future neural experiments: because the fixed view-angle condition was not run in the mouse studies, we cannot repeat this test in existing empirical data. Note that this pattern is not present in the scalar RPE, indicating that the loss function during training eliminates task-irrelevant features from the aggregated but not the individual RPEs (**Figure 3.12**).

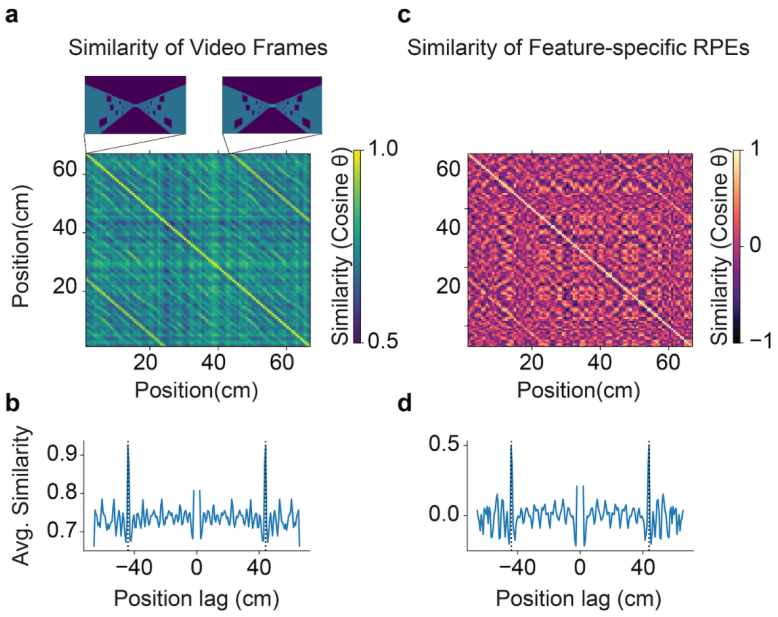


Figure 3.3: Feature-specific RPEs reflected incidental high-dimensional visual inputs. (a) Similarity matrices of the video frames, which measure the similarity between pairs of video frames (quantified by the cosine of the angle between them when flattened to vectors) across different position combinations. The off-diagonal bands correspond to the wall-pattern repetitions (see video frames for position 0cm and 43cm at insets above). **(b)** Average similarity

as a function of distance between frames, indicating that the average similarity peaked at the same position lag (43 cm) for video frames. **(c-d)** same as **(a-b)**, but with feature-specific RPEs.

Cue responses are consistent with feature-specific RPEs

While our feature-specific RPE model implies idiosyncratic and even task-irrelevant tuning in individual DA neurons, it also makes a fundamental prediction about the nature of these responses. In particular, responses to individual features represent not generic sensory or motor responses but feature-specific components of RPE. What in practice this means for a given unit depends both on what is its input feature ϕ_i, and also what other task features the brain represents (since net reward prediction and thus RPE are jointly determined by multiple cues). In general, we would expect that units that appear to respond to some feature (such as contralateral cues) do not simply reflect sensory responses (the presence of the cue) but rather should be further modulated by the component of RPE elicited by the feature. This could be particularly evident when considering the response averaged over units selective for a feature.

To test this hypothesis, we performed a new analysis, by subdividing cue-related responses (which are largely side-selective in both the model and neural data, **Figure 3.2g, j**) to determine if they were, in fact, additionally sensitive to the prediction error associated with a cue on the preferred side. For this, we distinguished these cues as *confirmatory* – those that appear when their side has already had more cues than the other and therefore (due to the monotonic psychometric curve, **Figure 3.2c**) are associated with an increase in the probability that the final choice will be correct and rewarded, i.e. positive RPE – vs *disconfirmatory* – cues whose side has had fewer towers so

far and therefore imply a decreased probability of reward (**Figure 3.4a**). As expected, when we considered the population of cue-onset responding feature-specific RPE units from the deep RL agent, these responses were stronger, on average, for confirmatory than disconfirmatory cues (**Figure 3.4b**), reflecting the component of RPE associated with the cue. We next reanalyzed our previous DA recordings based on this same insight. Consistent with the hypothesis that these cue responses reflect partial RPEs for those cues, the responses of cue-selective DA neurons were much stronger for confirmatory than disconfirmatory cues (**Figure 3.4c**). This implies that the heterogenous cue selectivity in DA neurons is indeed consistent with a cue-specific RPE. Importantly, the fact that these cue-responsive neurons are overwhelmingly selective for contralateral cues implies that these responses, combined across hemispheres, simultaneously represent two separate components of a vector of multidimensional feature-specific RPEs (as opposed to a classic scalar RPE). We confirm this by further separating the neural response between neurons recorded from each hemisphere (**Figure 3.13a,b**). As in the model (**Figure 3.13c,d**), responses to cues contralateral to each side both distinguish confirmatory from disconfirmatory cues.

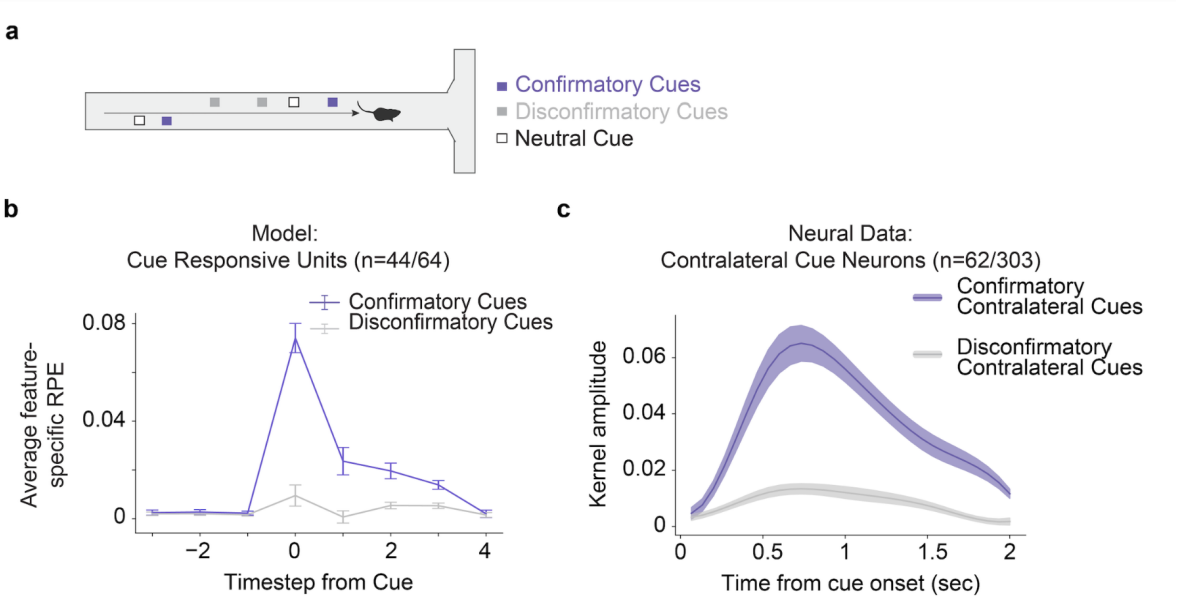


Figure 3.4: Cue responses in model and DAergic neurons reflected RPEs with respect to cues, rather than simply their presence. **(a)** Example trial illustrating confirmatory cues (purple), defined as cues that appear on the side with more evidence shown thus far, and disconfirmatory cues (gray), which are cues appearing on the side with less evidence shown thus far. Neutral cues (white) occur when there has been the same amount of evidence shown on both sides. **(b)** Average response of feature-specific RPE units modulated by cues (N = 44/64) to confirmatory (purple) and disconfirmatory cues (gray). Error bars indicate ±1 s.e.m. **(c)** Average responses of the contralateral cue responsive DA neurons for confirmatory (purple) and disconfirmatory (gray) contralateral cues. Colored fringes represent ±1 s.e.m. for kernel amplitudes (N = 62 neurons, subset of cue responsive neurons from **Figure 3.2j** that were modulated by contralateral cues only).

Uniform responses to reward at outcome period

In addition to explaining heterogeneity during the navigation and decision-making period, the feature-specific RPE model also explained the homogeneity of the neural responses to reward during the outcome period. Reflecting the standard properties of an RPE, the simulated scalar RPE (averaged over units) responded more for rewarded than unrewarded trials (**Figure 3.5a**). Since this aspect of the response ultimately arises (in Equation 1) from a scalar reward input, it is highly consistent across the units (**Figure 3.5b**), matching the neural data from our experiment (**Figure 3.5c**) and the widely reported reward sensitivity of DAergic units.

Equation 1 also implies a subtler prediction about the feature-specific RPEs, which is that although the modulation by reward is largely uniform across units, the modulation of this outcome response by the reward's predictability should be much more variable. This is because this latter modulation arises from the value terms in Equation 1, which are distributed across features (i.e., value is represented as a vector over features). In this task, reward expectation can be operationalized based on the absolute difference in tower counts, which is a measure of trial difficulty (predicting the actual chance of success as shown in **Figure 3.2c**), That is, when a reward occurs following a more difficult discrimination, the agent will have expected that reward with lower probability (**Figure 3.2d**). Accordingly, the simulated scalar RPE was larger for rewards on hard than easy trials (**Figure 3.5d**; **Figure 3.14**). However, when broken down unit-by-unit, although the median of individual units in the feature-specific RPE model was consistent with the scalar RPE ($P < 0.05$ two-sided Wilcoxon signed rank test for N=64), the size and direction of this effect varied widely across units (**Figure 3.5e**). A similar finding emerged from the neural data: while on average the reward response across reward-responsive population was modulated

as expected by expectation, there was high variability in the direction and extent of modulation across units (**Figure 3.5f**).

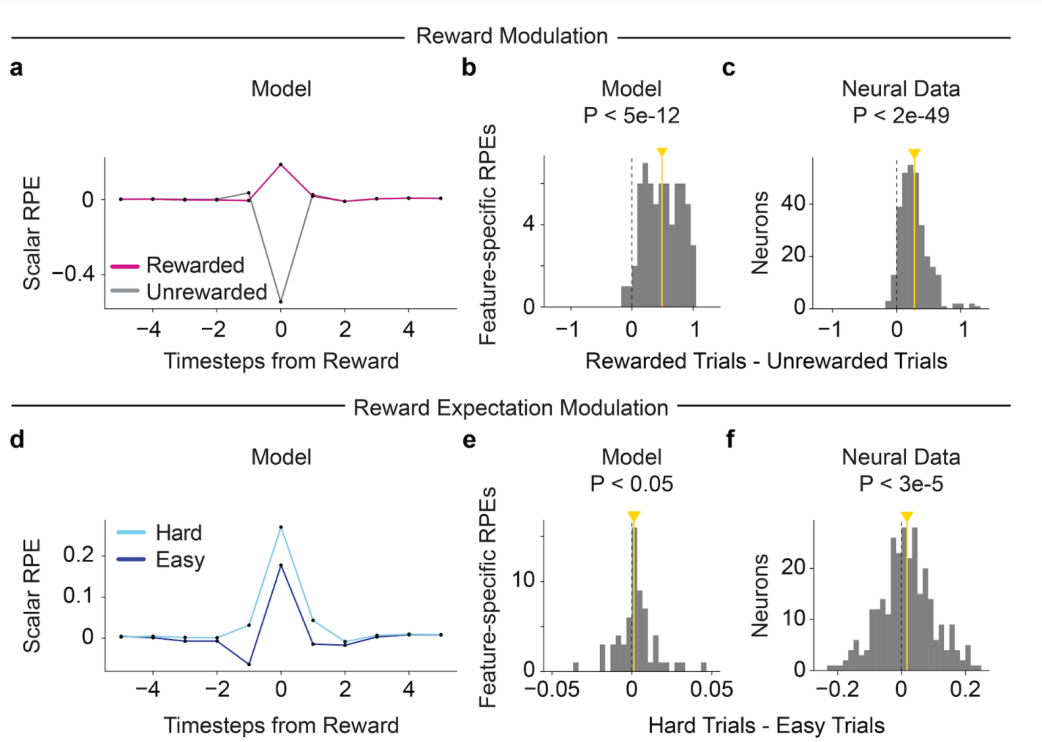


Figure 3.5: Feature-specific RPE units and DAergic neurons consistently respond to reward, but show heterogeneous modulation by reward expectation. (a) Scalar RPE (sum of feature-specific RPEs units) time-locked at reward time for rewarded (magenta) minus unrewarded trials (gray). **(b)** Histogram of feature-specific RPE units' response to reward minus omission at reward time ($P < 5e\text{-}12$ for two sided Wilcoxon signed rank test, $N = 64$). Yellow line indicates median. **(c)** Same as **(b)**, but with all imaged DA neurons(Engelhard et al. 2019), using averaged activity for the first 2 seconds after reward delivery, baseline corrected by subtracting the average activity 1 second before reward delivery ($P < 2e\text{-}49$ for two sided Wilcoxon signed rank test, $N = 303$). **(d)** Scalar RPE time-locked at reward time for rewarded trials, split by hard (light blue) and easy (dark blue) trials, defined as trials in the bottom or top tercile of trial

difficulty, measured by the absolute value of the difference between towers presented on either side. **(e)** Histogram of feature-specific RPE units' responses to difficult minus easy rewarded trials at reward time ($P < 0.05$ for two sided Wilcoxon signed rank test, $N = 64$) with median indicated by yellow line. There is an outlier data point at 0.29 for a feature-specific RPE unit showing strong reward expectation modulation. **(f)** Same as **(e)** but with the averaged activity of DA neurons(Engelhard et al. 2019) (Engelhard et al. 2019) corrected by their baseline activity 1 second before reward delivery ($P < 3e\text{-}5$ for two sided Wilcoxon signed rank test, $N = 303$).

The successor representation model cannot account for the reward-predictive aspects of VTA DA heterogeneity

Another candidate account of VTA DA heterogeneity is the successor representation (SR) model (**Figure 3.6a**) (Gardner, Schoenbaum, and Gershman 2018), which is an example of an outcome-specific PE model (**Figure 3.1b**). It posits that DA heterogeneity can arise from the prediction of different sensory features in different parallel circuits, rather than (or in addition to) prediction of reward. We simulated a set of SR error signals, by training a successor representation over the features produced by the DeepRL network trained on the task and then examining sensory prediction errors over these predictions. Similar to the neural data and to the feature-specific RPE model, the SR model produced heterogeneous responses during cue period (**Figure 3.6b-d**). However, these modulations were purely sensory rather than value-driven, and thus did not replicate the neural data. For example, responses to disconfirmatory cues were stronger than confirmatory cues (**Figure 3.6e**, opposite to the data and the feature-specific RPE model), because disconfirmatory cues are on average more surprising given the task statistics. Similarly, and contrary to the hallmark feature of VTA DA neurons, the SR model's outcome-period responses

are not consistently modulated by reward, showing no difference for averaged sensory PE during rewarded and unrewarded trials (**Figure 3.6f, h**) and also no consistent modulation by reward expectation (**Figure 3.6g, i**).

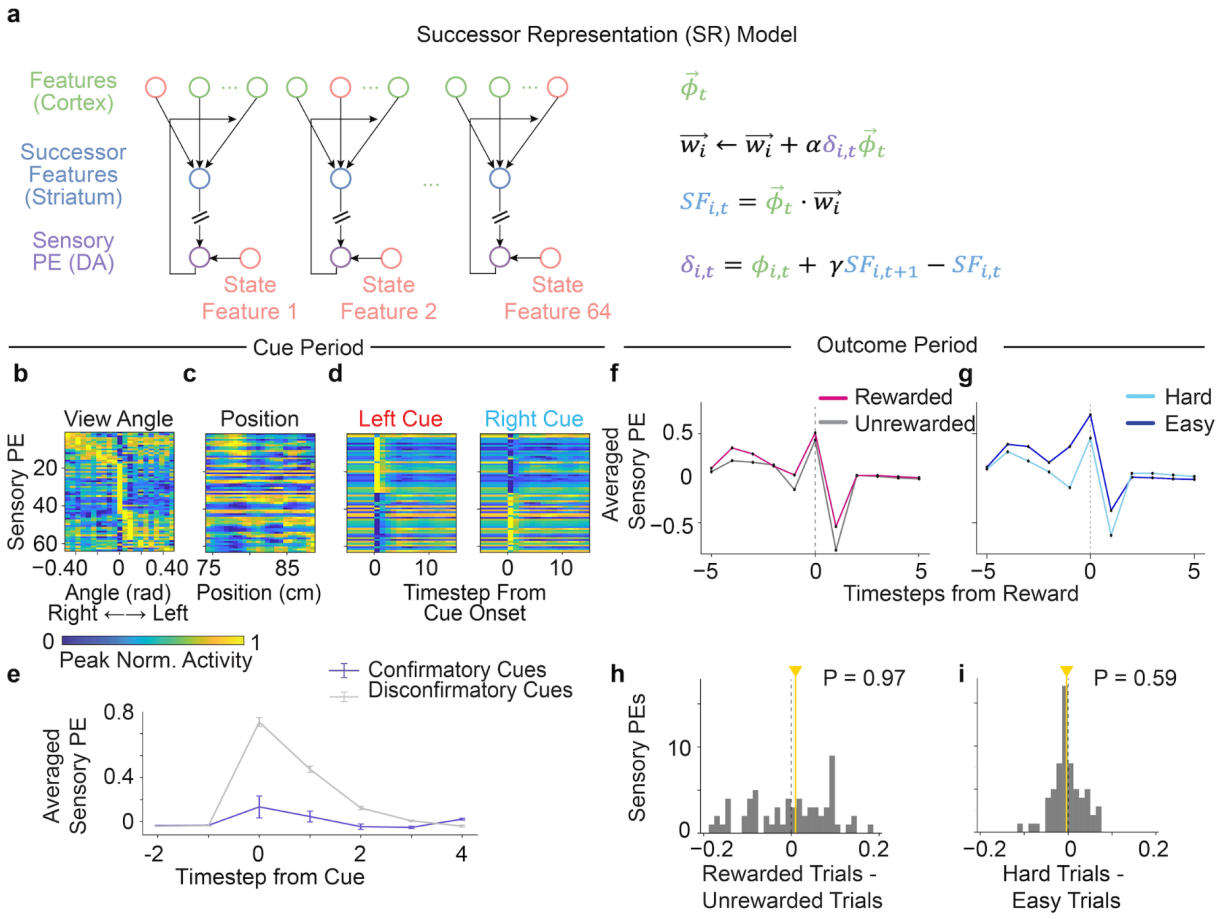


Figure 3.6: Sensory prediction errors from the successor representation model are heterogeneous during the cue period, but are not responsive to confirmatory cues and reward. **(a)** Schematic of the successor representation model to produce sensory prediction errors. The model was trained on the same state features as the feature-specific RPE model (green), with a vector of successor features (blue) and vector of sensory PEs (purple) with respect to each of the 64 features. **(b)** Sensory PEs plotted with respect to view angle. Each row corresponds to a unit's peak-normalized response to view angle averaged across trials. **(c-d)** Same as **(b)** but for **(c)**

position and **(d)** left (red) and right (blue) cues. **(e)** Sensory PE averaged across units in response to confirmatory (purple) and disconfirmatory cues (gray) shows a stronger response for disconfirmatory cues. Error bars indicate ±1 s.e.m. **(f)** Sensory PE averaged across units split by rewarded (magenta) and unrewarded (gray) trials (at the time of reward. **(g)** Same as **(f)** but for easy trials (dark blue) and hard trials (light blue) instead, defined as in **Figure 3.5d**. **(h)** Histogram of the difference in sensory PE units' responses at reward time for rewarded minus unrewarded trials, with the yellow line indicating the median (P = 0.97 for two sided Wilcoxon signed rank test, N = 64). **(i)** Same as **(h)** but for easy versus hard trials. Only rewarded trials are included (P = 0.59 for two sided Wilcoxon signed rank test N = 64).

Distributional RL cannot account for VTA DA heterogeneity during the cue period

Another prominent example of an outcome-specific PE model is the distributional RL model (**Figure 3.7a**) (Dabney et al. 2020), which posits that parallel channels predict different expectiles of the distribution of expected future reward, i.e. different weightings of better vs. worse outcomes. This model has been shown to predict subtle variability in cue and outcome DA responses in classical conditioning (Dabney et al. 2020), but it is unclear to what extent it might also capture the broader range of variability in more complex tasks like the one studied here. We simulated expectile PE signals based on the features produced by the DeepRL network trained on the task. This model did capture many of the reward-related aspects of the neural data, including stronger responses to confirmatory over disconfirmatory cues (**Figure 3.7e**) and uniform response to reward (**Figure 3.7f,h**). However, a subtle aspect of the reward response that was not captured by this model was the heterogeneity in the modulation of the reward response by expectation (**Figure 3.7g,i**).

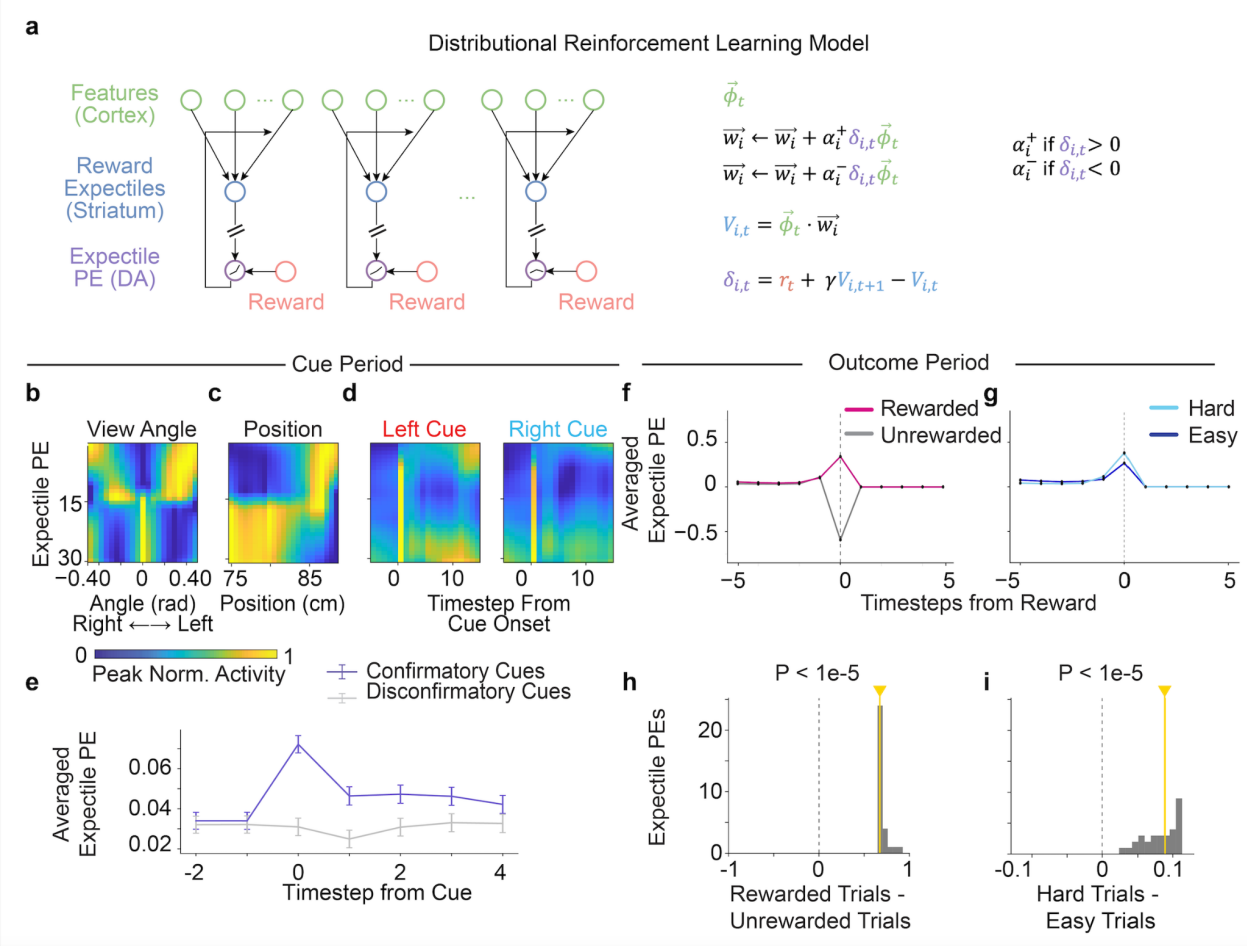


Figure 3.7: Expectile prediction errors from a distributional RL model are not heterogeneous during the cue period. **(a)** Schematic for the distributional reinforcement learning model which produces expectile prediction errors. The model uses the same state features as the feature-specific RPE model (green), with all the features input into 31 channels. These channels represent different expectiles of value (evenly spaced between 0 and 1), and 31 corresponding expectile prediction errors, which differentially weight positive and negative information. **(b)** Expectile PE activity plotted with respect to view angle, peak normalized for each unit and sorted in the original distributional channel order. **(c-d)** Same as **(b)** but with **(c)** position and **(d)** left (red) versus right (blue) cues. **(e)** Expectile PEs averaged across units, responding to confirmatory (purple) and disconfirmatory (gray) cues. Error bars indicate ±1 s.e.m. **(f)** Expectile PEs averaged across

units time-locked to reward, split by rewarded (magenta) versus unrewarded (gray) trials. **(g)** Same as **(f)** but split by easy trials (dark blue) and hard trials (light blue). Only rewarded trials are included, and trial difficulty is defined the same as **Figure 3.5d**. **(h)** Histogram of expectile PEs at reward time, showing the difference in their response for rewarded and unrewarded trials ($P < 1e\text{-}5$ for two-sided Wilcoxon signed rank test, $N = 31$ units). Median response shown with the yellow line. **(i)** Same as **(h)**, but for trial difficulty, taking the difference between responses for hard and easy rewarded trials ($P < 1e\text{-}5$ for two-sided Wilcoxon signed rank test, $N = 31$ units).

Feature-specific action prediction errors explain DA heterogeneity in SNc

So far we have shown that the feature-specific PE model captures, and outcome-specific PE models generally fail to match, heterogeneity of DA responses within VTA. This is consistent with our expectation that outcome-specific PEs are likely more relevant to heterogeneity between rather than within a DA subpopulation. To what extent does each class of model explain data from other tasks and other DA populations? Recent work has shown that DA neurons in the substantia nigra pars compacta (Parker et al. 2016; R. S. Lee et al. 2019; da Silva et al. 2018; Howe and Dombeck 2016) (SNc) and the substantia nigra pars lateralis (Greenstreet et al. 2022) (SNl) can exhibit qualitatively different response properties than those in the VTA, and that these differences may relate to their striatal projection target. In particular, some SNc/SNl DA neurons are tuned to actions, for example displaying a transient increase in firing shortly before action onset followed by a longer depression throughout action execution (Jin and Costa 2010; Dodson et al. 2016; Howe and Dombeck 2016; Coddington and Dudman 2018; da Silva et al. 2018). Interestingly, as a general rule, action selectivity in DA responses in SNc tends also to be accompanied by reduced response

to reward (Howe and Dombeck 2016; Parker et al. 2016; da Silva et al. 2018; Greenstreet et al. 2022). For example, Parker et al (Parker et al. 2016) recorded DA terminals in dorsomedial striatum (likely arising from SNc) during a reversal learning lever-choice task (**Figure 3.8a**) and showed that they responded robustly to contralateral movements (e.g., press of the lever contralateral to the recording site), but (in contrast to classic RPEs observed in nucleus accumbens, NAc, a VTA target) showed little modulation to reward consumption or reward-predicting cues.

Can outcome-specific PE or feature-specific RPE models explain the apparent tradeoff, across regions, between movement and reward selectivity? The action prediction error model (Greenstreet et al. 2022; Bogacz 2020; Lindsey and Litwin-Kumar 2022) is a type of outcome-specific prediction error model that offers a potential explanation. In this framework, there are parallel circuits that each predict different movements (e.g. left vs right lever press), instead of reward (**Figure 3.8b**, see Methods).

We compared the results of this model and our feature-specific RPE model (**Figure 3.8c**) in simulations of the Parker et al (Parker et al. 2016) probabilistic reversal task using a hand-designed "one-hot" state/feature space (**Figure 3.15a**, see Methods). We examined simulated DA responses from both models during choice and outcome, specifically focusing on the DA neurons from the left and right subcircuits in the APE model and the DA neurons associated with the left and right choice features in the feature-specific RPE model (**Figure 3.8d**). At choice, these make roughly similar predictions: the DA neurons fire more when their preferred action is made. At outcome, however, their predictions differ. Due to the r_t/N term in **Equation 3.1**, the feature-specific RPE DA neurons emit a fixed-magnitude, positive response to reward, whereas the APE

DA neurons do not. As such, while the feature-specific RPE model can produce DA neurons tuned to different movements, it does not explain the tradeoff between action selectivity and reward selectivity. The decoupling of actions from rewards suggests that for understanding why SNc-DMS responses differ from VTA-NAc ones, the principles of outcome-specific PE apply.

However, even within the movement-selective population in the SNc, there is additional functional heterogeneity. For example, Jin and Costa (Jin and Costa 2010) recorded individual DA neurons during a task in which mice had to sequentially press a lever multiple times for reward, and reported most DA neurons had specific preferred presses (e.g., the first or last presses within a sequence). Differential press-specific tuning is difficult to explain with a classic APE model. Since later presses are predicted by earlier presses, tuning for a standard APE unit should move backwards in time to the start of the sequence the same way it classically does in the scalar RPE model. Instead, we would expect this heterogeneity to reflect feature-specific computation within each APE circuit (**Figure 3.8f**).

To assess whether a feature-specific APE model could account for the results in Jin and Costa (Jin and Costa 2010), we simulated a simplified version of their task (**Figure 3.8e**) by augmenting the choice task from **Figure 3.8a** with a requirement of 3 presses of the lever to receive the reward (see Methods). To create predictions for DA responses on this task, we used both the standard APE model and a feature-specific APE model (**Figure 3.8f**) that functions like the feature-specific PE model but uses the state features to produce a set of prediction errors for (e.g.) left presses as outcomes instead of rewards (see Methods). For simplicity, we only analyze the left-selective circuits within each of these, but we note for completeness that they both include parallel left- and right-selective circuits.

For both models, the state representation was a one-hot encoding of the press count (**Figure 3.15b**). In the standard APE model, this predictably resulted in a response that peaked at the first press for all neurons and did not produce specific press selectivity (**Figure 3.8g, right**). In contrast, the feature-specific APE model showed uniform selectivity for press counts across all DA neurons (**Figure 3.8g, left**), with each DA neuron reflecting the tuning of its upstream state neuron. Thus, the feature-specific APE model recapitulates the experimentally reported press-selectivity across the population (Jin and Costa 2010). Furthermore, each feature-specific DA neuron exhibited a transient decrease in activity after its preferred press due to the feature-difference term in Equation 1, which is a property that has been observed in SNc DA neurons (Jin and Costa 2010; Dodson et al. 2016; Howe and Dombeck 2016; Coddington and Dudman 2018; da Silva et al. 2018).

One prediction of our model is that DA heterogeneity should be directly tied to the upstream cortical/striatal state features. For example, Jin and Costa (Jin and Costa 2010) observed that from day 1 to day 12 of training, the proportion of first-press and last-press neurons increased in tandem for both SNc DA neurons and striatal MSNs, with first-press increasing the most. To the extent the population of DA responses is indeed coupled to the upstream code, our model predicts that these changes imply a remodeling of the upstream population (simulated in **Figure 3.8h** by changing the state representation in our model of the task so that the first and last press are overrepresented; as expected, this induces the corresponding representational shift at the DA level). Indeed, this is consistent with evidence (including from Jin and Costa (Jin and Costa 2010) and also from other groups and tasks) that overtraining of this sort is also accompanied by a "chunking" of corticostriatal representation of the action sequence to "chunks" so that the overall

sequence becomes identified with its starting and concluding actions (Jog et al. 1999; Graybiel 2008).

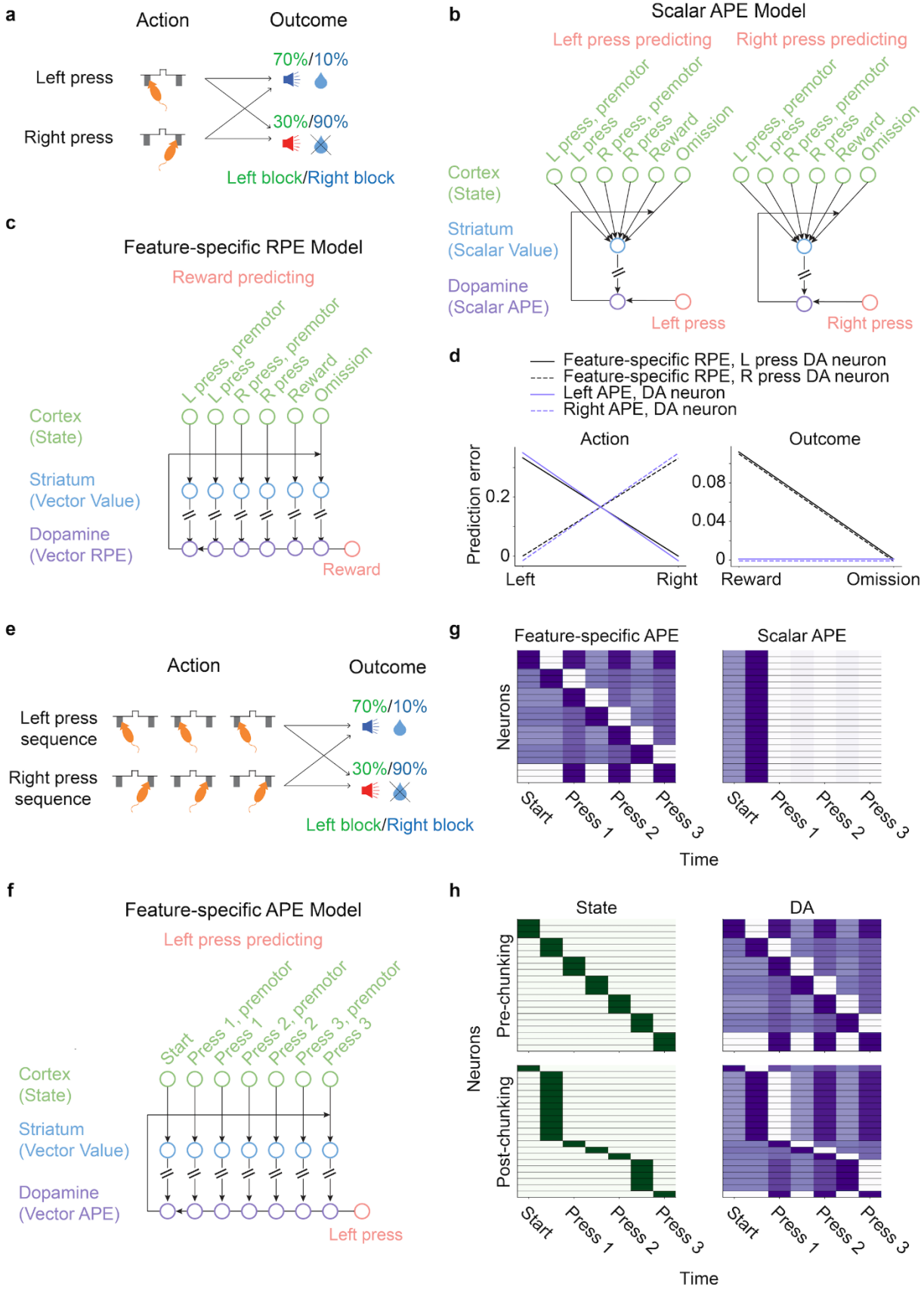

a
Action
Outcome
Left press
Right press
70%/10%
30%/90%
Left block/Right block
b
Scalar APE Model
Left press predicting
Right press predicting
L press, premotor
L press
R press, premotor
R press
Reward
Omission
Cortex (State)
Striatum (Scalar Value)
Dopamine (Scalar APE)
Left press
Right press
c
Feature-specific RPE Model
Reward predicting
Cortex (State)
Striatum (Vector Value)
Dopamine (Vector RPE)
Reward
d
Feature-specific RPE, L press DA neuron
Feature-specific RPE, R press DA neuron
Left APE, DA neuron
Right APE, DA neuron
Action
Outcome
Prediction error
0.2
0
0.08
0.04
0
Left
Right
Reward
Omission
e
Action
Outcome
Left press sequence
Right press sequence
70%/10%
30%/90%
Left block/Right block
f
Feature-specific APE Model
Left press predicting
Start
Press 1, premotor
Press 1
Press 2, premotor
Press 2
Press 3, premotor
Press 3
Cortex (State)
Striatum (Vector Value)
Dopamine (Vector APE)
Left press
g
Feature-specific APE
Scalar APE
Neurons
Start
Press 1
Press 2
Press 3
Time
h
State
DA
Neurons
Pre-chunking
Post-chunking
Start
Press 1
Press 2
Press 3
Time

Figure 3.8: Feature-specific action prediction errors explain movement-related heterogeneity in SNc DA neurons. **(a)** Schematic for the Parker et al(Parker et al. 2016) reversal task. Mice were offered a choice between two levers, each of which provided reward with some probability. During a left block (green), the left lever rewarded with probability 0.7 and the right lever rewarded with probability 0.1. During a right block (blue), this contingency was reversed. Identity of the current block (left or right) was unsignaled to the mice and swapped with a pseudorandom schedule. **(b)** Schematic of the scalar APE model. The model was trained with features encoding a one-hot representation of the state (green), scalar action value (blue), and scalar APE (purple). The left schematic displays the left-preferring subcircuit and the right schematic displays the right-preferring subcircuit. **(c)** Schematic of the feature-specific RPE model used on the Parker et al reversal learning task(Parker et al. 2016). Input features were the same as the scalar APE model in **(b)**. The value layer (blue) and the RPE layer (purple) were feature-specific. **(d)** Dopaminergic predictions from the scalar APE model (purple) and the feature-specific RPE model (black) during action execution (left) and outcome (right) in the reversal learning task(Parker et al. 2016). **(e)** Schematic for a simplified version of the Jin and Costa task(Jin and Costa 2010). Mice are offered a choice between two levers, each of which provided a reward with some probability after being pressed three times. Reward contingencies are the same as **(a)**. **(f)** Schematic for the left subcircuit of the feature-specific APE model used on the Jin and Costa(Jin and Costa 2010) task. The model was trained with features encoding the state (green), vector action value (blue), and feature-specific APEs. **(g)** Activity of the DA neurons in the feature-specific (left) and scalar (right) APE models during the task. Neurons are sorted by peak onset. Black horizontal lines separate the neuronal activity profiles. **(h)** Activity of the state (green) and DA neurons (purple) in the feature-

specific APE model before (top) and after (bottom) chunking. Black horizontal lines separate the neuronal activity profiles.

3.4 Discussion

Here we propose a new theory which helps to reconcile recent empirical reports of DA response heterogeneity to the classic idea of DA neurons as encoding RPEs. Most previous work has considered "outcome-specific" PEs that describe heterogeneity between parallel circuits for predicting different outcomes. We introduce a complementary "feature-specific" model, where heterogeneity arises from the high-dimensional state input to each such circuit. We show how this model, but not alternative models based on outcome-specific PEs, produces heterogeneous responses to task variables, but relatively uniform responses to reward, recapitulating recent empirical work (Engelhard et al. 2019; Kremer et al. 2020). We also test the model's prediction that heterogeneous DA responses are not simply responses to sensory and behavioral features of the task, but instead reflect components of the RPE with respect to a subset of the features. Finally, we show how the principles of outcome- and feature-specific PEs combine to explain different aspects of movement-related heterogeneity in SNc DA neurons (Jin and Costa 2010).

Feature-specific vs outcome-specific PEs: Physiology

Our central observation is that TD models imply two distinct candidate sources of heterogeneity across DA units: that arising from the target outcome (e.g., reward) and that arising from the state input (e.g., cues and other task features). Most previous theories have explained DAergic heterogeneity by positing multiple distinct error signals, each specialized for learning to predict a different outcome (Dabney et al. 2020; Gardner, Schoenbaum, and Gershman 2018; Kurth-Nelson

and Redish 2009; Daw, Courville, and Touretzky 2006, 2003; Rao 2010; Lak et al. 2017; Lloyd and Dayan 2016; Tano, Dayan, and Pouget 2020; Ribas-Fernandes et al. 2011; Botvinick, Niv, and Barto 2009; Gershman, Pesaran, and Daw 2009; Jiang and Litwin-Kumar 2021; Daw, Kakade, and Dayan 2002). For instance, a family of error signals could be used to learn to predict different outcomes such as rewards versus punishments (Daw, Kakade, and Dayan 2002), to learn the rewards associated with different actions or effectors (Gershman, Pesaran, and Daw 2009; Greenstreet et al. 2022; Bogacz 2020; Lindsey and Litwin-Kumar 2022), to predict rewards at different temporal scales (Kurth-Nelson and Redish 2009; Wei, Mohebi, and Berke 2021), or to track different goals and subgoals in a hierarchical task (Ribas-Fernandes et al. 2011; Botvinick, Niv, and Barto 2009; Frank and Badre 2012). In contrast, the basic insight of our new model is that within any such outcome-predicting circuit, nonuniform input projections from an arbitrary population code for state can give rise to diverse patterns of simultaneous, multiplexed responses to different nonreward task events. Collectively, these responses constitute a population code over feature-specific PEs in an otherwise standard TD learning setting aiming (in each circuit) to predict each single, scalar outcome like reward.

Though both classes of accounts are quite general, we can still differentiate their respective contributions to observed heterogeneity in a number of ways. The first is that they imply different relationships between heterogeneous responses at outcome (to rewards, omissions, or punishments (Dabney et al. 2020; H. Matsumoto et al. 2016; de Jong et al. 2019; Menegas et al. 2018; Lerner et al. 2015; M. Matsumoto and Hikosaka 2009; Lammel et al. 2012; Verharen, Zhu, and Lammel 2020; Cai et al. 2020; Syed et al. 2016; Howe and Dombeck 2016; O'Doherty et al. 2004)) versus heterogeneous responses to other task events, including stimuli and movements (Engelhard et al.

2019; Kremer et al. 2020; Moss et al. 2021; Zolin et al. 2021; da Silva et al. 2018; Syed et al. 2016; Howe and Dombeck 2016; Parker et al. 2016; R. S. Lee et al. 2019; O'Doherty et al. 2004; Barter et al. 2015; Saunders et al. 2018; Jin and Costa 2010). Outcome-specific PEs imply that variation between neurons in their responses to stimuli ultimately arises from those neurons specializing in different outcomes. For this reason, these models fail to capture the coexistence in VTA DA between heterogeneous responses to task variables alongside a more uniform main effect of reward (Engelhard et al. 2019). This is instead a hallmark of a feature-specific RPE, in which different neurons represent different stimulus-specific components of a PE for a common reward. Our simulations show that different outcome-specific models can either produce heterogenous coding of cues and task variables by predicting many outcomes, but then fail to respond to reward (**Figure 3.6**, successor representation) or instead maintain the overall reward effect by focusing on predicting a narrower range of reward-related outcomes, but with little variability in responses to cues and other task variables (**Figure 3.7**, distributional RL).

A related prediction of the feature-specific RPE is that DA encoding of task features, in general, actually reflects components of RPE with respect to each feature, rather than strictly the main effect of each feature itself. For instance, in our data and in the feature-specific RPE model, different DA neurons are tuned for different cues (left vs. right), but each population further distinguishes confirmatory from disconfirmatory cues (**Figure 3.4b,c**), which differ in their reward consequences. In contrast, different SR neurons respond to the cues themselves, not their reward associations (**Figure 3.6d,e**). An important aspect of the data is that the responses are tied to individual features – consistent with the feature-specific model's unique claim that different neurons represent RPE components for left cues and right cues separately. This stands in contrast

to a distributional RL model (**Figure 3.7a**) and also to classic scalar TD (**Figure 3.1a**), both of which also predict that cue responses should be modulated by their reward associations (e.g., confirmatory vs. disconfirmatory **Figure 3.7e**), but wrongly predict that all neurons should respond similarly to left vs. right cues, which do not have distinct reward associations (**Figure 3.7d**).

Importantly, because both frameworks are so general – different PEs can be defined for any vector of outcomes, and also for any vector of features – we have here emphasized qualitative predictions that aim to capture what is common to each entire family (Palminteri, Wyart, and Koechlin 2017). We have also, for concreteness, used a deep network model to learn a detailed and reasonably realistic high-dimensional feature set with which to simulate the model and expose its key behaviors. However, this model involves a number of simplifications (for instance, it does not vary running speed or other kinematics, and does not carry any information from trial to trial) which mean that it cannot in principle capture some types of responses reported in the neural data (Engelhard et al. 2019), and is also likely to quantitatively differ with respect to others (e.g. head direction, which is more constrained in simulation). One aspect of the response that is notably different in simulation is position-related ramps. While these do arise in simulation – and demonstrate how feature-specific PEs can, for appropriate features, capture even temporally extended responses – they tend in the model to occur only over shorter distances than in DA recordings, possibly reflecting how the model features were built (unrealistically) via backpropagation through time. In any case, the phenomenon of positional ramping in DA is under active investigation, and a full explanation is likely to involve a number of additional

considerations (e.g. uncertainty) not included in the current model (H. R. Kim et al. 2020; Mikhael et al. 2022; Gershman 2014).

Feature-specific vs outcome-specific PEs: Anatomy

An important dimension predicted to distinguish feature- from outcome-specific heterogeneity is anatomical. DA neurons reporting different outcome-specific PEs should be linked to their targets in a closed loop, meaning this form of heterogeneity should be found across projection-defined DA populations. In contrast, heterogeneity arising from the feature-specific PE model is nested within each outcome-specific circuit, meaning it should exist even within a projection-defined population of DA neurons. In fact, feature-specific decompositions of a PE will best retain the strengths of the scalar model when units representing different aspects of PE have overlapping targets, because this would allow PEs for different aspects of state to mix into a common update.

Together, these considerations lead to the overall prediction that the characteristics of outcome- vs feature-specific PE variation should predominate, respectively, between vs within different projection-defined DA populations. Formally investigating this prediction (e.g., by systematically comparing cue and outcome responding across different projection-defined DA populations) is the key open empirical test of our framework. However, existing data are generally consistent with the hypothesis that the sorts of heterogeneity associated with outcome-specific differences tend to arise between distinct DA nuclei or target regions (Tsutsui-Kimura et al. 2020; Parker et al. 2016; Avvisati et al. 2022), in contrast to the hallmarks of feature-specific PEs we study here, which occur within VTA (Engelhard et al. 2019; Kremer et al. 2020) and also within SNc (Jin and Costa 2010). Thus, in SNc and DMS, compared to VTA and NAc, movement responses tend to emerge

while reward responses decline (a combination consistent with outcome- rather than feature-specific variation; **Figure 3.8**) (Parker et al. 2016; Howe and Dombeck 2016; Jin and Costa 2010; Saunders et al. 2018; Barter et al. 2015); DA neurons responding to threat rather than reward have been reported in projections from lateral Snc to the tail of the striatum (Menegas et al. 2018, 2017); and recent data suggest that the differential responsivity to RPE magnitudes that is characteristic of distributional RL also emerges across DA projections to distinct striatal subregions (Tsutsui-Kimura et al. 2020; Avvisati et al. 2022). A related point is that, since the feature-specific PE model predicts that averaging over features will recover the original scalar PE, the types of variation it envisions will tend to be washed out by methods like photometry that involve averaging over many neurons.

While more anatomical data is needed to fully clarify the scope of the feature-specific and outcome-specific models, existing data regarding inputs to DA neurons, as well as DA innervation to striatum, already provides important information. Consistent with our feature-specific model, inputs to DA neurons (especially from striatum), which in our model represent the state representation, are topographic, and plausibly more precisely so than the ascending DA projection (Lau, Monteiro, and Paton 2017; Pan, Mao, and Dudman 2010; Cox and Witten 2019; Alexander, DeLong, and Strick 1986). These afferent neurons are known to reflect the same sorts of heterogeneity with respect to task features as we consider at the DA level (Jin and Costa 2010; Parker et al. 2022; Tsujimoto, Genovesio, and Wise 2011; Rueda-Orozco and Robbe 2015; K. Choi et al. 2023; Robbe 2018; N. Matsumoto et al. 2001), consistent with our framework in which the set of DA units that collectively target some striatal domain are assumed to vary in their feature inputs (e.g. **Figure 3.8h**). Regarding DA projections, individual DA neurons branch extensively

to innervate large areas (~1 mm^3) of striatum (Matsuda et al. 2009), and volumetric propagation of released DA to extrasynaptic DA receptors may further blur DA's effect (Arbuthnott and Wickens 2007; Caillé, Dumartin, and Bloch 1996; Rice and Patel 2015; Gonon et al. 2000). This suggests a limited number of sufficiently non-overlapping outcome-specific PE circuits are possible, although how many remains unknown. Recent work partly moderates these concerns by emphasizing that DA signaling may be somewhat more precise than the initial observations suggest, in that only a minority of DA varicosities have active release sites, and DA reuptake can limit diffusion (Liu and Kaeser 2019; Liu et al. 2018; Beyene et al. 2019).

Credit assignment and outcome- vs feature-specific PEs

Previous outcome-specific PE models have typically been motivated as enabling enhanced computational function: additional PEs compute additional quantities (Daw, Kakade, and Dayan 2002; Lloyd and Dayan 2016; Gardner, Schoenbaum, and Gershman 2018; Dabney et al. 2020; Akiti et al. 2022; Rao 2010; Lak et al. 2017; Daw, Courville, and Touretzky 2003; Kurth-Nelson and Redish 2009; Gershman, Pesaran, and Daw 2009). In contrast, the motivation for our model is a pair of more practical issues with this picture. The first is the need for distinct closed-loop circuits for each PE, which likely constrains the scale of heterogeneity that can play this role. The second is the requirement that information about many widely dispersed state features be brought together as inputs for value prediction and RPE. We suggest that these two issues are related: while it is likely inevitable that inhomogeneous state feature input drives response variation within a projection-defined DA population, the same overlapping ascending projections that constrain the outcome-specific PEs also help to promote convergence across these feature-specific PEs.

Such convergence, in turn, addresses a central computational function: credit assignment over multidimensional feature spaces. Normatively, the scalar RPE arises algebraically from gradient descent and also from statistical inference, when a prediction is made from a weighted sum of features (i.e., the delta rule is a special case of both backpropagation and also Bayes' rule (Rumelhart, Hinton, and Williams 1986; Kalman 1960; Widrow and Hoff 1960)). Functionally, the effect of such a common RPE is to enable stimuli to compete with one another to explain observed rewards, each stimulus learning uniquely to explain what cannot be explained by the others. In Pavlovian conditioning, this insight goes back to the classic model of Rescorla and Wagner, who pointed out that such competitive credit assignment via a shared RPE is observable behaviorally in the blocking effect: an animal will not learn that some stimulus predicts reward, if the reward is already predicted on the basis of another stimulus (Rescorla 1972; Kamin 1967).

These considerations provide a normative, computational rationale why – to whatever extent RPE heterogeneity arises from feature inputs rather than outcome targets – it is advantageous that they mix (as in **Figure 3.1c**) rather than remain separate (as in **Figure 3.1b**), and thereby reveal an unrecognized virtue of the diffuse ascending anatomy. They also offer, via the blocking effect, a behavioral test of this prediction. If stimuli instead have separate errors, PE-driven learning still works for each stimulus individually (this was the basis of conditioning models predominant prior to Rescorla-Wagner (Estes 1950; Pearce and Hall 1980)), but they will not engage in competitive credit assignment. This means our model degrades gracefully if, in practice, mixing over the feature-specific RPEs is imperfect (e.g., if ascending projections do not perfectly overlap). To whatever extent different DA subpopulations represent RPEs specific to different features, but without anatomically overlapping projections, their respective cues will fail to block one another

behaviorally. For instance, if VTA DA responses to towers are largely lateralized in opposite hemispheres in Pavlovian conditioning, as they are in the Engelhard et al (Engelhard et al. 2019) task, then (although the data lack projection tracing) their targets are unlikely to overlap, and such a failure of blocking could be expected between left and right towers.

Although cue competition via a common error signal offers a well-justified and simple approach to credit assignment, if cues are instead separated in a parallel-circuit architecture like the outcome-specific PE models, then credit assignment requires an additional computational approach. Such models, collectively known as "mixtures of experts," use separate prediction modules associated with different subsets of stimulus features and represented in different closed-loop channels (Doya et al. 2002; Singh 1992; Hamid, Frank, and Moore 2021). Mixture of expert schemes offer more individualized control, to "divide and conquer" and focus learning on context-relevant dimensions. However, unlike cue competition through a scalar error signal, in this case credit assignment requires an external arbitration mechanism to determine which modules to rely on in any particular context (Doya et al. 2002; Singh 1992; Hamid, Frank, and Moore 2021). One recent model (Hamid, Frank, and Moore 2021) uses this type of architecture to suggest a different set of mechanisms addressing some aspects of the problems we discuss here. In particular, this work suggests that even within a broadcast error signal, the timing of a scalar DA response (arriving earlier or later due to wave-like response propagation) can target the signal to different modules. The study also suggests arbitration principles for externally assigning credit to different modules by routing waves depending on circumstances, though the detailed neural basis of such arbitration remains to be explored. Overall, to what extent the brain relies on either credit assignment mechanism remains a question, but the present framework clarifies how these issues

could be examined via behavioral (e.g. blocking), projection-tracing and neurophysiological studies.

Population codes for state

Although much RL modeling in neuroscience assumes a simple state representation that is hand-constructed by the modeler (Montague, Dayan, and Sejnowski 1996; Schultz, Dayan, and Montague 1997), in general, an effective state representation depends on the task, and the brain must learn or construct it autonomously as part of solving the full RL problem. How it does this is arguably the major open question in these models. Indeed, in AI, recent progress on this problem (notably using deep neural networks) has been the main innovation fueling impressive advances scaling up otherwise standard RL algorithms to solve realistic, high-dimensional tasks like video games (Mnih et al. 2015; Silver et al. 2016). In psychology and neuroscience also, there have been a number of recent theoretical hypotheses addressing how the brain might build states, such as the successor representation and latent-state inference models (Gershman, Blei, and Niv 2010; Gershman, Norman, and Niv 2015; Courville, Daw, and Touretzky 2005; Russek et al. 2017; Stachenfeld, Botvinick, and Gershman 2017; Niv 2019). But there exist relatively few experimental results to assess or constrain these ideas, for instance because learning behavior alone is relatively uninformative about state, whereas in the brain it is unclear which neural representations directly play this role for RL (e.g., grid cells vs. place cells for spatial tasks (Gustafson and Daw 2011; Stachenfeld, Botvinick, and Gershman 2017)).

A main consequence of the feature-specific PE model is that, if it is correct, then the heterogeneous DAergic population itself gives a new experimental window, from the RL system's perspective,

into the brain's population code over state features. While the model itself is agnostic to the feature set used, the various DA responses should in any case reflect it. This builds on previous work showing that even a scalar DAergic TD error signal can be revealing about the upstream state that drives it (Wilson et al. 2014; Y. K. Takahashi and Schoenbaum 2016; Y. K. Takahashi et al. 2011), but on the new theory, the feature-specific vector DA code much more directly reflects the upstream distributed code for state. Thus, the theory offers a general framework to reason quantitatively about population codes for state. This should enable new experiments and data analyses to infer the brain's specific state representation from neural recordings and in particular to test ideas about how it is built: how it changes across different tasks and as tasks are acquired.

3.5 Supplementary Materials

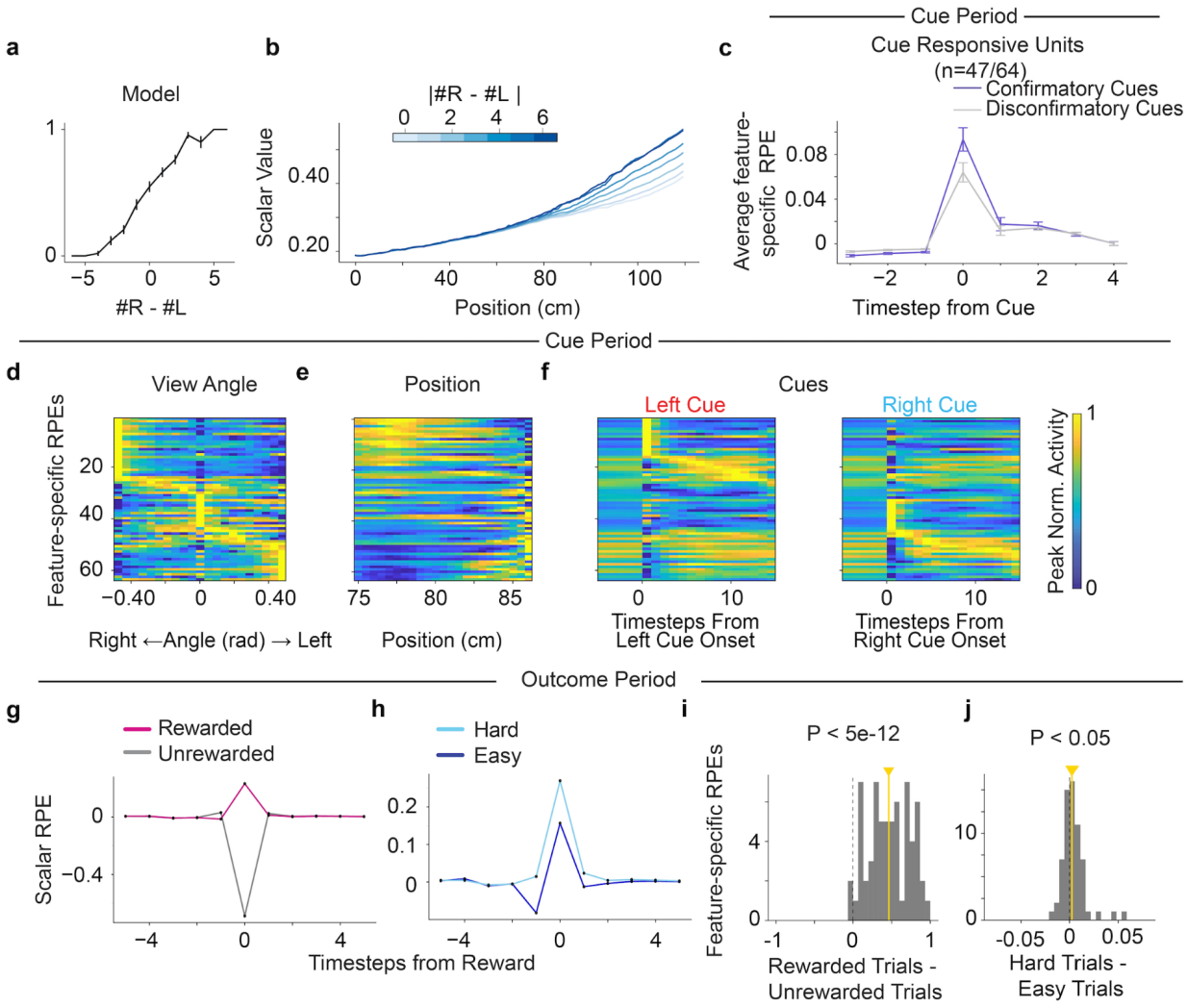


Figure 3.9: Deep reinforcement learning network retrained with the same task and a different seed. **(a)** Psychometric curve showing the retrained agent's performance. **(b)** Retrained agent's scalar value during the cue period decreased as a function of trial difficulty (defined as the absolute tower difference, blue gradient). **(c)** Retrained agent's feature-specific RPE units' response to confirmatory (purple) and disconfirmatory (gray) cues. Response is averaged across cue-sensitive units only (N = 47/64). **(d)** Retrained agent's feature-specific RPE units averaged across trials plotted with respect to view angle. Each row represents one unit's peak normalized response to the view angle. **(e-f)** Same as **(d)** but for the agent's **(e)** position and **(f)** left (red)

and right (blue) cues. **(g)** Retrained agent's scalar RPE time-locked at reward time for rewarded (magenta) and unrewarded trials (gray). **(h)** Same as **(g)** but for rewarded trials with different trial difficulties, with hard trials (light blue) and easy trials (dark blue) defined like in **Figure 3.5d**. **(i)** Histogram of retrained agent's feature-specific RPE units' response to reward minus omission at reward time ($P < 5e\text{-}12$ for two sided Wilcoxon signed rank test, $N = 64$). Yellow line indicates median. **(j)** Same as **(i)** but for rewarded trials plotted against trial difficulty ($P < 0.05$ for two sided Wilcoxon signed rank test, $N = 64$). In **(j)**, there is an outlier data point at 0.16 for a feature-specific RPE unit showing strong reward expectation modulation.

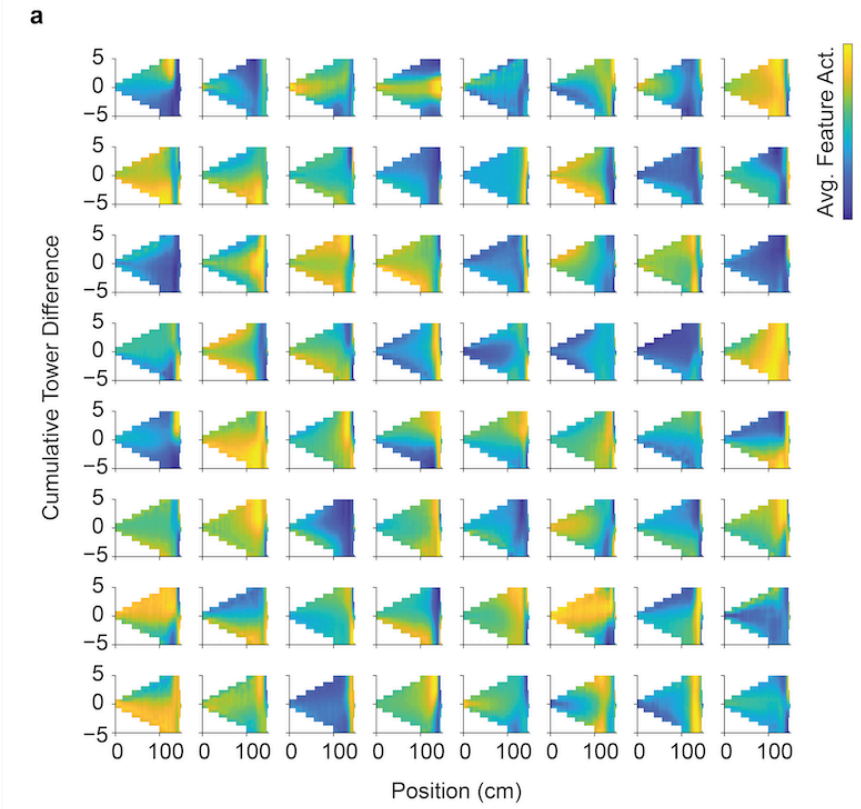


Figure 3.10: Tuning of 64 LSTM feature units to position and evidence. (a) Each panel shows an individual feature unit and how it is tuned to the agent's position in the maze and the cumulative tower difference at that position.

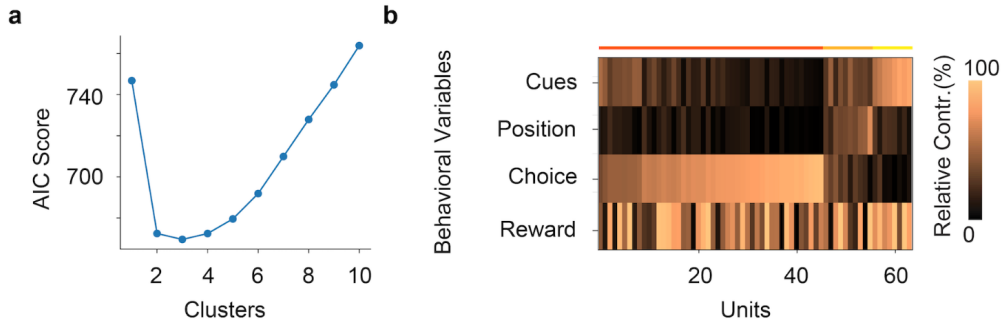


Figure 3.11: Feature-specific RPEs clustered to behavioral features of the task to match Engelhard et al (Engelhard et al. 2019) **clustering analysis. (a)** Optimal number of clusters was 3, selected by minimizing the AIC scores for models with different numbers of clusters. **(b)** Relative contribution of the behavioral features including cues, position, choice and reward response for the 64 units sorted based on the highest probability belonging to the cluster, with colored lines on top indicating the cluster identity. Relative contribution is defined as the percentage of the explained variance for the partial model not including the variable versus the full model.

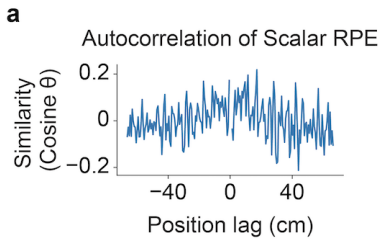


Figure 3.12: Scalar RPE signal does not reflect the incidental high-dimensional features. (a) Scalar RPE signal autocorrelated across time (similarity defined as the cosine of position-lagged scalar RPE responses) does not show peaks at position lag 43 cm for the wall-pattern repetition location.

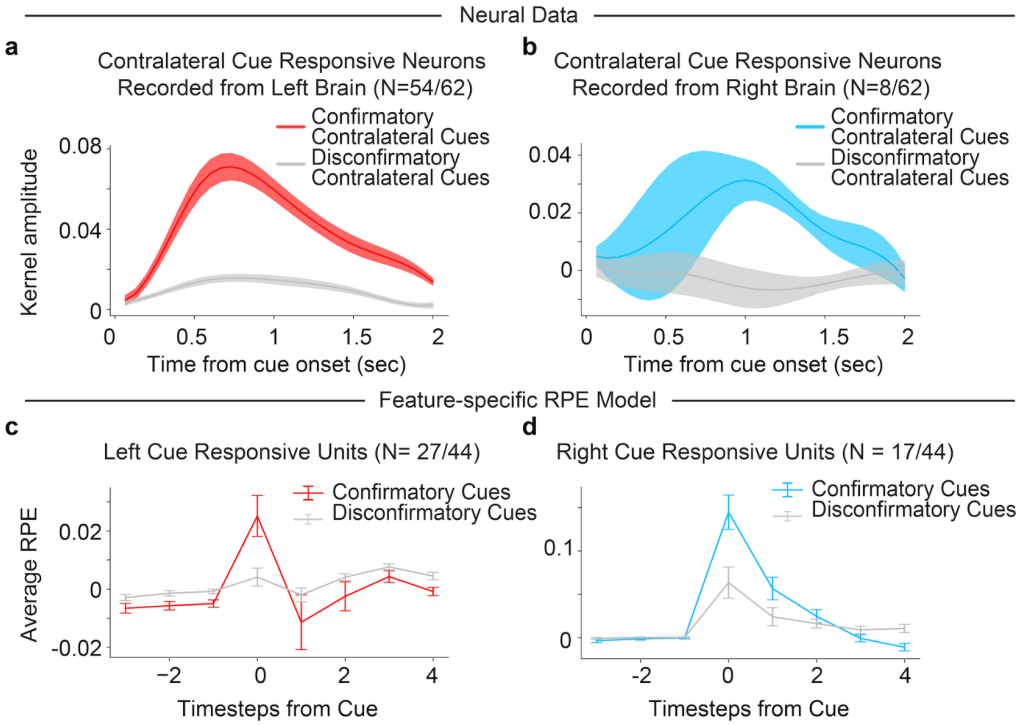


Figure 3.13: Cue responses in model and DAergic neurons represent a vector code of lateralized response and RPE. **(a)** Average responses of the contralateral cue responsive DA neurons (Engelhard et al. 2019) only recorded from the left hemisphere (N= 54/62 neurons, subset of contralateral cue responsive neurons from **Figure 3.4c**) for confirmatory (red) and disconfirmatory (gray) contralateral cues. Colored fringes represent ±1 s.e.m. for kernel amplitudes. **(b)** Same as **(a)** but for DA neurons recorded on the right hemisphere (N = 8/62) responding to confirmatory (blue) and disconfirmatory (gray) confirmatory cues. **(c-d)** Same as **(a-b)**, but for feature-specific RPE model units responding to **(c)** left cues specifically (N = 27/44, subset of the cue responsive neurons from **Figure 3.4b** that were modulated by left cues only) and **(d)** right cues specifically (N = 16/44). Error bars indicate ±1 s.e.m.

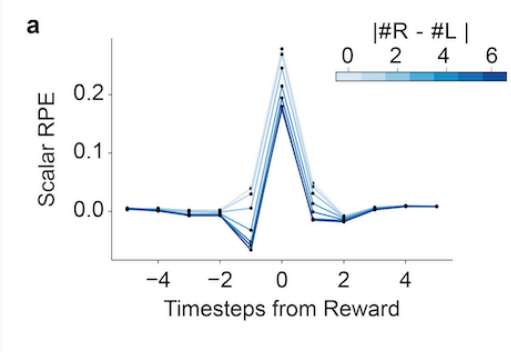


Figure 3.14: Scalar RPE shows fine-grained reward expectation modulation. (a) Scalar RPE's response modulated by reward expectation given by the difficulty of the task, defined as the absolute value of the final tower difference (blue gradient) of the trial.

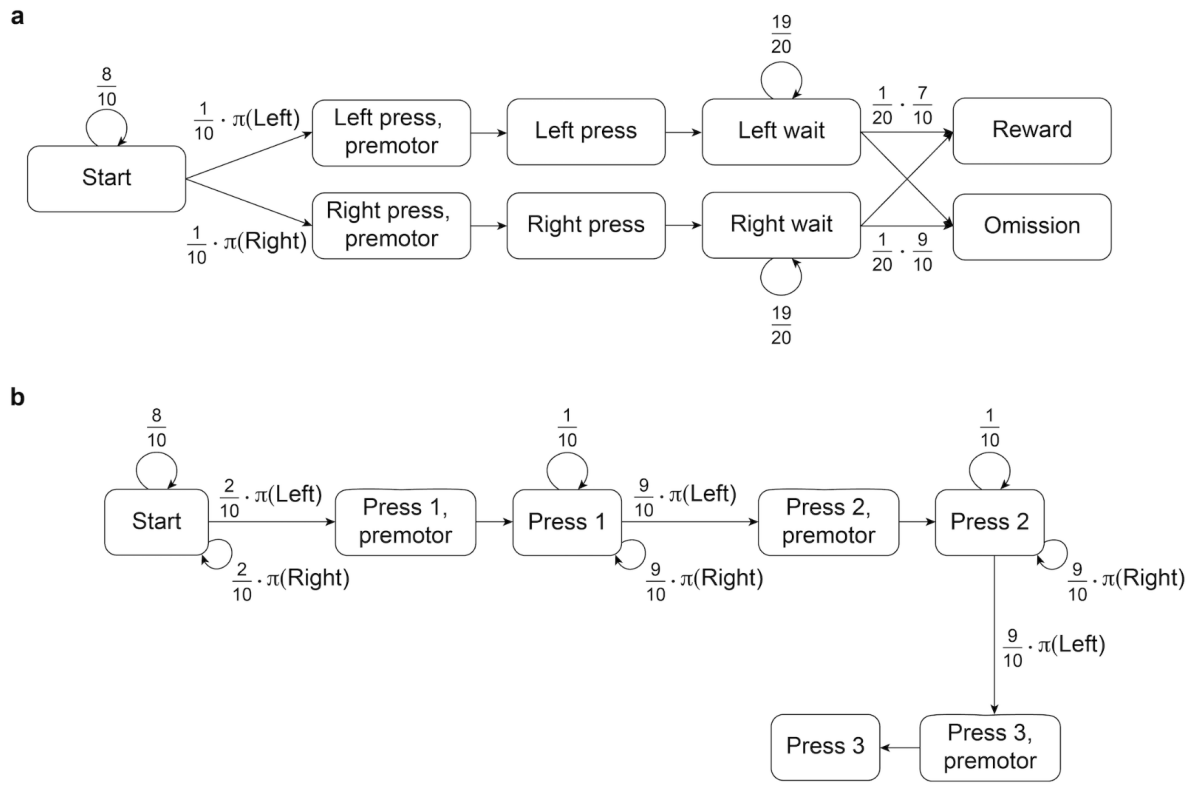


Figure 3.15: State diagrams for simulations of the Parker et al (Parker et al. 2016) **and Jin and Costa** (Jin and Costa 2010) **tasks in Fig. 8. (a)** State diagram for the Parker et al (Parker et al. 2016) task simulation. **(b)** State diagram for the Jin and Costa (Jin and Costa 2010) task simulation. In both panels, arrows indicate probabilistic transitions between states, with probabilities described by the arrow labels. Unlabeled arrows denote transitions with the remaining probability to make the total sum to 1. $\pi(x)$ refers to the probability of executing action x under the agent's behavioral policy π.

3.6 Acknowledgements

We thank Alvaro Luna and Juan Lopez for their help with the VR software system, Will Dabney for discussion on the distributional RL model and providing the imputation function for the distributional RL model, Michelle Lee and Erin Grant for help with training the deep RL network, Peter Dayan, Ari Kahn, and Lindsey Brown for comments on this work, and additionally the BrainCOGs team and the entire Daw and Witten labs for their help. This work was supported by an NSF GRFP (RSL), 1K99MH122657 (BE), NIH R01 DA047869 (IBW), U19 NS104648-01 (IBW), ARO W911NF-16-1-0474 (NDD), ARO W911NF1710554 (IBW), Brain Research Foundation (IBW), Simons Collaboration on the Global Brain (IBW), and the New York Stem Cell Foundation (IBW).

Chapter 4

Understanding Value-free Action Prediction Error in the Brain

4.1 Introduction

Dopamine (DA) neurons in the basal ganglia primarily serve to modulate movement and reward-motivated behavior (Niv et al. 2007; Berke 2018). Generally, these two distinct functions are segregated by the dorsal versus ventral pathways, with dorsal striatum projecting DA neurons reflecting movement and ventral striatum projecting DA neurons reflecting reward (Saunders et al. 2018; Parker et al. 2016). In Chapter 2, we affirm this topographical functionality by showing how DMS-projecting DA neurons directly signal movement rather than some reward prediction error with respect to movement (R. S. Lee et al. 2019). In Chapter 3, we revisit this dataset by seeing if we can simulate the functional segregation by using our feature-specific prediction error model, but use movement instead of reward to drive prediction error (R. S. Lee et al. 2023).

In this chapter, we want to revisit the dataset in Chapter 2 once more to see if it can show that DA is not just signaling movement, but a value-free movement prediction error. Previous work relating DA to movement has shown that DA can signal movement initiation, invigoration, or choices (Wise 2004; R. S. Lee et al. 2019; Dodson et al. 2016; Howe and Dombeck 2016; da Silva et al. 2018; Barter et al. 2015; Syed et al. 2016; Jin and Costa 2010). However, a recent report takes this a step further, showing how DA can reflect value-free prediction errors related to the action chosen by the rodent (Greenstreet et al. 2022). Greenstreet et al. recorded tail-striatum (TS) projecting DA neurons while mice perform an auditory discrimination task: mice get rewarded as they press the lever associated with the appropriate auditory stimuli (Greenstreet et al. 2022). During training, the DA populations do not respond to reward at outcome time, but

instead to contralateral choices when presented with the auditory cue (Greenstreet et al. 2022). They argue that these are in fact prediction errors by first showing that DA responses were not sensitive to value-changes in the task, but changes to the state (i.e. when the auditory cue linked with the lever was changed to white noise) (Greenstreet et al. 2022). Next, they optogenetically stimulate the DA populations and show that this increases mice's preference for the contralateral lever (Greenstreet et al. 2022). These new results suggest that DA's role as a carrier of prediction error could be far richer than we previously thought.

While the empirical result for APE is fairly new, several theory and modeling papers have utilized a movement PE (Bogacz 2020; Lindsey and Litwin-Kumar 2022), one of which focused on how such a learning signal might be used to facilitate habitual behaviors (K. Miller, Shenhav, and Ludvig 2019). These models are particularly compelling because while extensive work has shown how striatal regions are related to complementary behaviors– specifically, dorsal medial striatum (DMS) for goal-directed behavior (i.e. learning some action-outcome contingency) (Yin et al. 2005; Gremel and Costa 2013) and dorsal lateral striatum (DLS) for habitual behavior (i.e. learning some stimulus-response association) (Yin, Knowlton, and Balleine 2004; Yin and Knowlton 2006; Tricomi, Balleine, and O'Doherty 2009)–, there have not been extensive studies showing how DA projecting to these regions might facilitate these behaviors. One possibility is that previous studies showing DA modulation to movement might be also signaling movement PE, and revisiting these results might help elucidate how DA can signal PE related to reward and movement together to help coordinate goal-directed and habitual behaviors.

In light of these empirical and theoretical results, we reinvestigate the Parker dataset from Chapter 2, which has DA recordings from VTA/SN that terminate in DMS and DLS while mice perform a probabilistic reversal learning task (Parker et al. 2016). Specifically, we hope to see whether or not we can see the corresponding PE signal in each of the striatal projections, thus showing how DA inputs to these striatal areas might reflect the goal-directed and habitual behavior with which their target striatal regions are associated. We test whether or not these neural signals might be reflecting these PE signals by adapting a regression technique from (Bayer and Glimcher 2005). Specifically, if DA is reflecting reward PE, regressing the signal against rewards from past trials should yield coefficients that are significantly positive or negative depending on the sign for how we define reward PE. Similarly, we can test if the signal also reflects action PE by regressing the signal against action history, and checking that the sign of the coefficients match how we define action PE.

4.2 Methods

Dataset

The dataset in this chapter is the same as the one used in Chapter 2, so most of the methods for the mice, the surgery, and data processing remain the same and will not be repeated in this chapter.

The only new data consist of the VTA/SN::DLS terminal recordings, which were unpublished both in the original paper (Parker et al. 2016) and the followup paper that makes up Chapter 2 (R. S. Lee et al. 2019). For the VTA/SN::DLS terminal recordings, Cre-dependent GCaMP6f

virus (AAV5-CAG-Flex-GCamp6f-WPRE-SV40; UPenn virus core, injected titer of 3.53 × 1012 pp per ml) was injected into VTA/SNc while fibers were placed in the DLS (M-L ± 2.1mm, A-P 0.74mm, D-V -2.9mm), with one recording area per mouse for a total of N = 7 recording sites. 2 of the mice were recorded bilaterally, with the second site in VTA/SN::DMS terminals. 6 of the mice were recorded from the left hemisphere, with 1 from the right hemisphere. All mice were male.

To summarize, there were 368.56 trials per mouse for VTA/SN::DMS terminal recordings and 1160.14 trials per mouse for VTA/SN::DLS terminal recordings. The VTA/SN::DMS terminal data consisted of 10108 total trials across 12 recording sites, and the VTA/SN::DLS terminal data consisted of 8121 total trials across 7 recording sites.

Prediction Error Regression

In **Figure 4.2,** we fitted a linear regression model for each mouse to determine if the DA signals reflected both reward and action prediction errors. We regressed the averaged DA signal (window of GCaMP6f recording for each striatal target shown on **Figure 4.2a,d** in yellow) against reward in the past four trials, current trial's choice, the choices in the past four trials, and an intercept. The VTA/SN::DMS DA data used a window of 9 timesteps (0.6 seconds) and the VTA/SN::DLS DA data used a tighter window of 6 timesteps (0.4 seconds) to account for the faster, sharper response to lever presentation (Wei, Mohebi, and Berke 2021). We fitted coefficients for each mouse, then averaged across all of the mice for a grand mean (N = 12 for VTA/SN::DMS, N = 7 for VTA/SN::DLS) with the error bars indicating s.e.m.. In order to

determine if the fitted coefficients are significantly positive or negative, we used a two-sided Wilcoxon signed rank test for each coefficient across the mice.

In **Figure 4.3**, we fitted a linear regression model with the same GCaMP6f windows from each striatal target, but we changed the dependent variables. Instead of previous trial's choices, we used an indicator variable for whether or not the previous trial's choices matched with the current trials' choice. Each previous trial's choice is compared with the current trial's choice, and the current trial's choice was also included as a dependent variable. The coefficients were treated the same way, with the two-sided Wilcoxon signed rank test performed for each variable.

4.3 Results

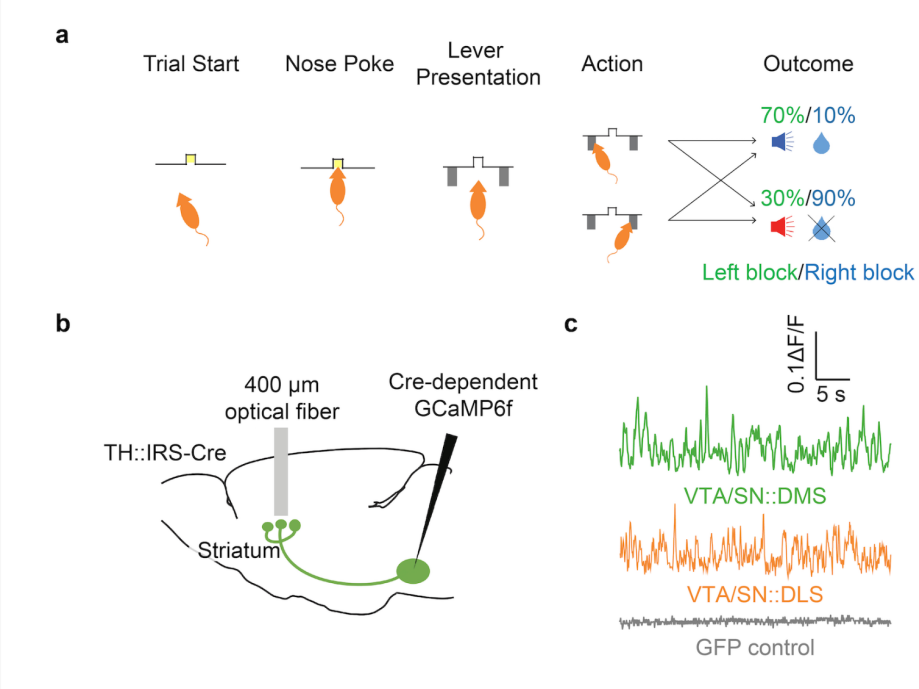


Figure 4.1: Mice learn a reversal learning task while GCaMP6f is recorded from two different DA striatal targets, VTA-SN::DMS terminals and VTA-SN::DLS terminals. **(a)** Reversal learning **(b)** Schematic of the surgery to set up the optogenetic recording, with the GCaMP6f recording originating from VTA/SN and terminating in different striatal targets. **(c)** Example GCaMP6f recordings from VTA/SN::DMS terminal (green) and VTA/SN::DLS terminal (orange). Bottom row (gray) is an example recording of a mouse expressing GFP control.

Mice performed a reversal learning task (**Figure 4.1a)** while we recorded from DA neurons with projections to dorsal medial striatum (DMS) and dorsal lateral striatum (DLS) (**Figure 4.1b-c**). In each trial, the center port was illuminated to prompt the mouse to nosepoke. After a short delay of 0-1 seconds, two levers were presented to the mouse on the left and right of the center port. The animal must choose one of the two levers, which yielded a high probability (70%) or low probability (10%) of reward. The mouse waited for another 0-1 seconds of delay, before a pure tone auditory stimulus indicated a sucrose reward in the center port (positive conditioned stimulus, i.e. CS+) or a white noise auditory stimulus indicated that there is no reward. The key is that the identity of the high probability lever was swapped after at least 10 trials plus some random number of trials, but not the change was not signaled to the mouse. Thus, the mouse must continuously seek out the high probability lever, switching between levers as they get more or less reward.

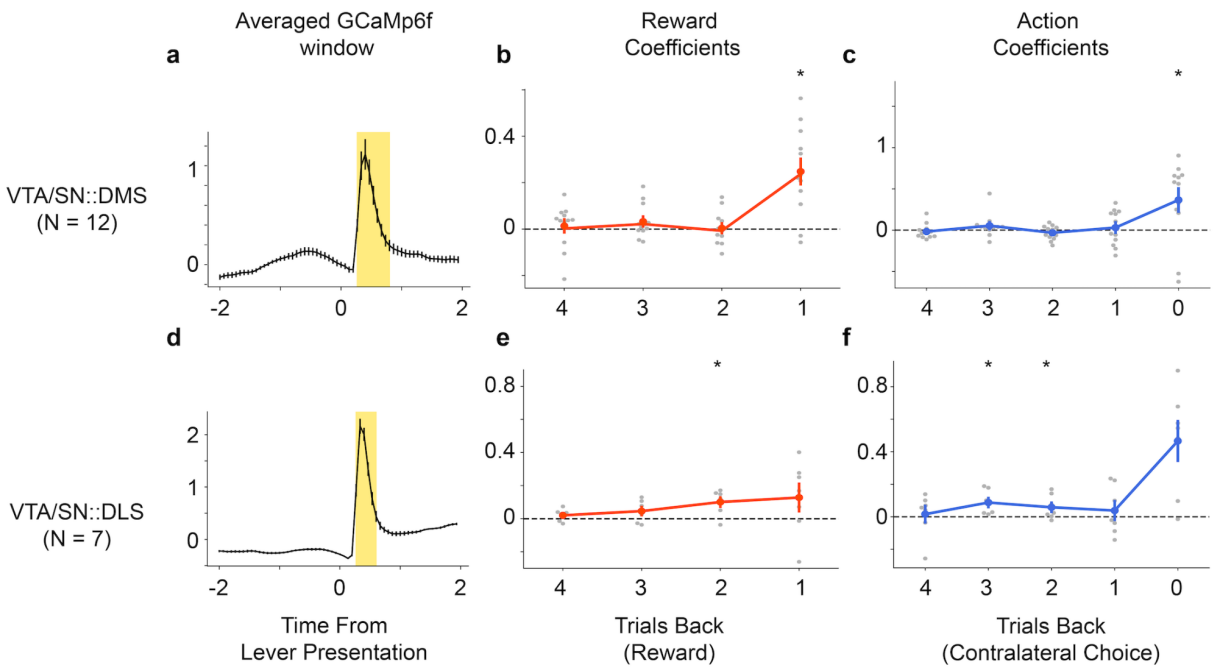


Figure 4.2: VTA/SN::DMS (N= 12) and VTA/SN::DLS (N= 7) terminal recordings averaged and regressed against previous rewards, current choice, and previous choices does not show a clear RPE or APE signal. **(a)** GCaMp6f recording from VTA/SN::DMS terminals timelocked to lever presentation, with the window used for the averaged GCaMP6f indicated in yellow (9 timesteps). **(b)** Regression coefficients for rewards from previous trials and **(c)** choices in current trial and previous trials. Each mouse''s regression coefficient is given as a gray dot. The averaged coefficient is plotted over for rewards (red) and choices (blue), and the error bars indicate s.e.m. (N = 12). Star above each averaged coefficient indicates if P value < 0.05 with a two-sided Wilcoxon signed rank test. **(d-f)** The same as **(a-c)** but for VTA/SN::DLS terminals. The window in **(d)** is 6 timesteps long and N = 7 for **(e-f)**.

Like in the Greenstreet paper (Greenstreet et al. 2022), we focus on the GCaMP6f signal right before the animal makes a decision, hence we take averaged GCaMP6f signals timelocked at

lever presentation (**Figure 4.2a,d**). In order to determine if we have a reward and action PE signal, we regressed averaged GCaMP6f signals against the rewards in previous trials, current choice, and choices in previous trials.

We can define the reward PE signal at lever presentation as:

$$\delta_t^{rew} = V_t \text{ [4.1]}$$

with V_t as the value of the chosen lever. This is derived by taking the TD error equation (**Equation 1.1**), but setting the reward as 0. Replacing V_t with the update step for TD error, we can now define value with respect to previous trials, or:

$$\begin{aligned}\delta_t^{rew} &= \alpha^{rew}[r_{t-1} - V_{t-1}] + V_{t-1} \\ &= \alpha^{rew} r_{t-1} + (1-\alpha^{rew}) V_{t-1} \\ &= \alpha^{rew} r_{t-1} + (1-\alpha^{rew})[\alpha^{rew} r_{t-2} + (1-\alpha^{rew}) V_{t-2}] \\ &= \alpha^{rew} r_{t-1} + (1-\alpha^{rew}) \alpha^{rew} r_{t-2} + \ldots (1-\alpha^{rew})^{k-2} \alpha^{rew} r_{t-k} \text{ [4.2]}\end{aligned}$$

with k defined as the number of trials back and α^{rew} as some learning rate. Thus, we expect that the PE signal should depend on the previous trial's reward positively, with previous trials' rewards' effect exponentially decreasing by a factor of $(1-\alpha^{rew})$.

Looking at the regressed neural data, both VTA/SN::DMS terminal and VTA/SN::DLS terminal recordings' coefficients somewhat matched with the expected pattern signaling reward

PE (**Figure 4.2b,e**). For the VTA/SN::DMS terminal DA responses, we see positive coefficients for previous trials' reward, with the two-trial back having an exponentially smaller positive coefficient but trials further back around the same levels (**Figure 4.2b**). Only the previous trial's response has a significant p-value, showing that the reward PE response is only speculative (**Figure 4.2b**). The VTA/SN::DLS terminal DA responses had a more consistently decreasing positive coefficients for previous trials, but did not show the exponential decrease of the coefficients on trials back and also only had a significant p-value on the two-trial back coefficient (**Figure 4.2e**). Overall, while the pattern of coefficients somewhat indicated a reward PE signal, not enough coefficients were significant nor cleanly matched the expected rate of decrease.

For the action PE signal, we can look to Miller et al. for the equations (K. Miller, Shenhav, and Ludvig 2019):

$$\delta_t^{act} = a_t - H_t \text{ [4.3]}$$

$$H_{t+1} = H_t + \alpha^{act} \delta_t \text{[4.4]}$$

Here, we have δ_t^{act} as the action PE, $a_t$as the choice in the current trial (coded as some indicator function for the action chosen), α^{act} as some learning rate, and H_t as the recency weighted average of how often a choice was taken previously. While other action PE signals are defined, they overlap in that they are still taking some difference between the actions chosen and the averaged actions taken so far (Bogacz 2020; Lindsey and Litwin-Kumar 2022; Greenstreet et al. 2022). If an action PE signal is present, we would expect to see positive modulation by the current choice, along with negative modulations on previous trial's choices.

This is because in **Equation 4.3**, we have a PE given as the difference between the choice taken minus the averaged choices of previous trials.

Both striatal targets' DA responses showed a positive modulation for the current trial's choice, but did not have negative coefficients for previous trials (**Figure 4.2c, f**); this suggests that these DA responses only indicate contralateral movement modulation and not a movement PE. Like in Chapter 2, the VTA/SN::DMS terminals had a significant response to contralateral movement (**Figure 4.2c**). However, VTA/SN::DLS terminals did not show a significant response to contralateral movement, only the contralateral choice two and three trials back. Generally, we cannot conclude that we see an action PE from either striatal targets, and thus we cannot see how DA might modulate goal-directed and habituated behaviors via their different striatal targets.

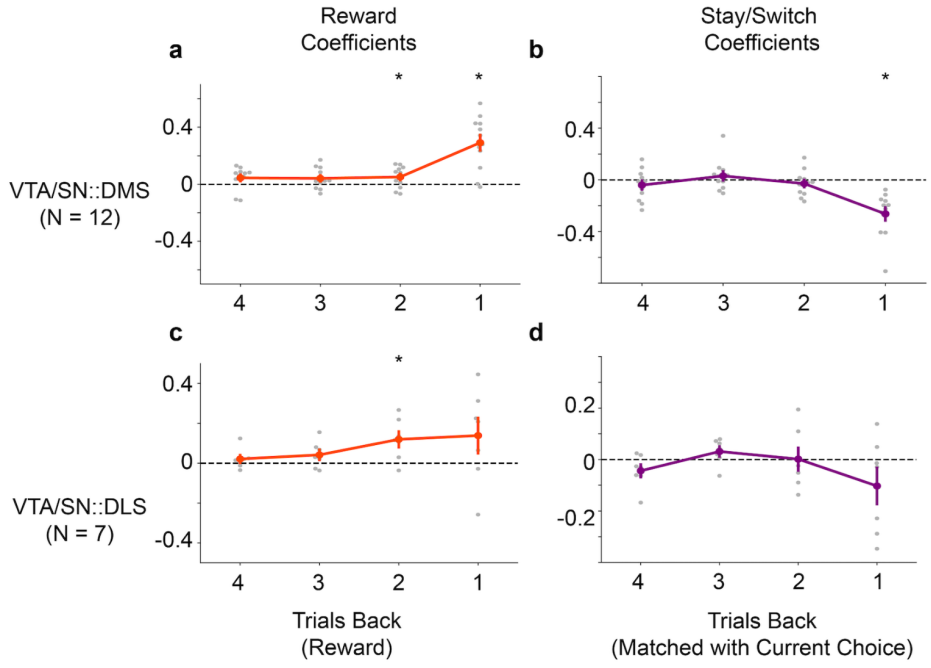


Figure 4.3: Regression coefficients for VTA/SN::DMS (N= 12) terminal recordings reflect both an RPE and identity-free APE signal, while coefficients for VTA/SN::DLS (N= 7)

terminal recordings did not. (a) Coefficients from regressing averaged VTA/SN::DMS terminal signal (window same as **Figure 4.2a**) for rewards from previous trials and **(b)** whether or not current trial's choice matched with the one on previous trials. Gray dots indicate individual mouse's coefficients, while the grand mean is plotted for rewards (red) and matched trials (purple). Error bars show s.e.m. (N = 12), and statistically significant (P-value < 0.05 with a two-sided Wilcoxon signed rank test) averaged coefficients are denoted with a star. **(c-d)** The same as **(a-b)** but for VTA/SN::DLS terminals. The averaged GCaMP6f signal uses the same window shown in **Figure 4.2d.**

However, we could consider an action PE that is not sensitive to the identity of the choice, but indicates if the current choice is different from previous choices (i.e. the habits). In this case, we would be regressing not against previous choices' identity, but whether or not the current choice matched with previous trial's choices. Such a signal is also described as an "action-surprise" signal (Lindsey and Litwin-Kumar 2022), which specifies that the PE is not action-specific and produces the same value if two distinct actions are equally surprising. Mathematically, we would modify **Equation 4.3** to be an L2 norm, effectively masking the choice's identity:

$$\delta_t^{act} = \| a_t - H_t \|^2 \quad [4.5]$$

In this case, we see that VTA/SN::DMS terminals GCaMP6f response do, in fact, encode both an reward PE and action PE signal. The previous trial's reward coefficients fit the exponentially decreasing trend and are significant for the first two previous trials (**Figure 4.3a**) as predicted by **Equation 4.2**. We also see a significantly negative coefficient (N = 12) for the previous trial's matched choice, indicating that the DA signal is stronger when the previous trial was surprising

from the choices so far (**Figure 4.3b**). While the two-trial back matched action is not significant, we do see a slightly negative coefficient (**Figure 4.3b**).

The VTA/SN::DLS terminal recordings, on the other hand, still did not show a significant reward or action PE signal. While the sign for the previous trial's reward and the previous trial's matched choices fitted roughly with the expected pattern, the coefficients were not significant under a two-sided Wilcoxon signed rank test (N = 7) (**Figure 4.3c-d**).

4.4 Discussion

As a followup to analyses and simulations in previous chapters, we revisit the Parker et al. dataset in order to see if we can reinterpret the contralateral preference of dorsal-projecting DA neurons as a value-free movement PE signal. By looking at both DMS and DLS-projecting DA populations, we aimed to link each of the projection target's functional role in goal-directed and habitual learning respectively to whether or not the DA signals reflected a movement PE signal, which has been previously proposed in theoretical models as a way to keep track of and maintain habits. While we did not find a clear signed movement PE in either projection target's DA populations, we did find statistically significant correlates of an unsigned action PE signal in DMS-projecting DA populations during lever presentation. These results do not support our original hypothesis that DLS-projecting DA neurons might reflect a stronger movement PE signal given its established role in forming and maintaining habitual behavior (Yin and Knowlton 2006; Yin, Knowlton, and Balleine 2004; Gremel and Costa 2013; Tricomi, Balleine, and O'Doherty 2009); however, it does show how the contralateral movement signals in the

DMS-projecting DA neurons could still be formulated as a PE rather than simply a movement signal.

We maintain that our results are preliminary and there are several considerations that we need to keep in mind with respect to the dataset and analyses. First, we would want more recording sites for the DLS-projecting DA neurons, especially since the dataset isn't counterbalanced between sides. Our results did show that the DLS-projecting DA neurons may reflect an action PE, but the coefficients were not significant. Perhaps with more samples, we could get clearer results. Second, the regression is done by taking the average GCaMP6f signal at lever presentation, rather than timestep by timestep. We picked the window based on the peak from the average response of all the mice, but examining the time between lever presentation to lever press across all trials shows that the delay can last as long as 10 seconds. This would affect how well the coefficients fit with respect to the neural data, given that the behavior is rather variable. Finally, we define the unsigned movement signal by whether or not the animal repeated the same choice in the current trial in previous trials. In any task, and in reversal learning tasks in particular, stay versus switch decisions can be correlated with other features of the task such as reward. Additional analysis would be needed to be able to clearly show that there is indeed an unsigned movement signal, including a more significant coefficient from two- and three- trial back matches.

Finally, if there does exist an action surprise signal, more work would need to be done to theorize why such a signal might exist and how it might be involved in facilitating reinforcement learning. Previous theoretical work that has implemented this signal used gradient methods;

thus, while the PE signal itself is agnostic to the actions chosen, the weight updates still use the gradient of the signal that includes the action chosen (Lindsey and Litwin-Kumar 2022). In our formulation, this surprise signal is scalar and not reward-modulated, and therefore could not serve as a teaching signal alone to keep track of choice history. Thus, instead of updating some habitual controller as in **Equation 4.4**, the action surprise signal could act as a salience signal, highlighting when the agent has chosen differently than before. This signal can then act as a gating mechanism, and the weight updates might need to be done through Hebbian learning instead.

Chapter 5

Conclusion

5.1 Summary of Results

In each chapter, I explored a puzzling set of dopamine data, and sought to build updated versions of classic TD learning to explain the results. In Chapter 2, I attempted to understand movement correlates in DMS-projecting DA populations by modeling the neural data and seeing if it might reflect a reward prediction error with respect to the contralateral movement. However, I could not find such a signal, only a direct movement signal. I revisit this dataset later on in Chapter 4, hoping to find a reward-irrelevant movement prediction error instead. While I could not find a signed movement prediction error response, I did find an unsigned action surprise signal. This action prediction error model is an updated version of classic TD learning, using movement as the motivating input rather than reward. These results remain preliminary, and more work would need to be done to actually show that such a signal exists and more theoretical work would be needed to understand why this signal would be useful.

In Chapter 3, I focused on a different set of DA data that showed VTA DA populations reflecting seemingly reward-irrelevant behavioral features before the animal receives the reward. I introduced the feature-specific prediction error model, which hypothesized that these heterogeneous results would be a reflection of the state features. I apply this model to the aforementioned heterogeneous VTA data, and also extend it to show how it might explain heterogeneous movement responses in SnC as well.

5.2 Revisiting Parker et al. Dataset

In my dissertation, we have modeled the Parker et al. dataset with multiple models, each with different conclusions. An important point of discussion is how we can interpret the data overall, given how I treat the data with a new lens each time.

In both Chapter 2 and Chapter 4, we see that DMS-projecting DA neurons reflected a movement signal directly (**Figure 2.3a-c, Figure 4.2c,f**). We have tried to interpret this movement signal as a prediction error in multiple ways every chapter: in Chapter 2, as a RPE with respect to contralateral movement (**Figure 2.2a**); in Chapter 3, as an action PE (**Figure 3.8b**) and a feature-specific action PE (**Figure 3.8c**); in Chapter 4, as an unsigned action PE, or action surprise, signal (**Equation 4.4**). Most of these attempts have failed when confronted with neural data. Looking at the trial-by-trial RPE signal in Chapter 2, we see that the DA signal could not be explained by a reward PE. Simulations in Chapter 3 showed that an APE model could explain how the contralateral sensitivity might arise (**Figure 3.8d**), but further examination in Chapter 4 showed that the neural data did not fit an action-specific APE model (**Figure 4.2c,f**).

Instead, the only promising result was when we fitted the neural data to an action-surprise APE signal that simply indicated if the current choice was different (or surprising) from the choices so far (**Figure 4.3b**). While more work is needed to confirm the results, this could offer a way movement and reward-motivated functions of DA can be explained under a unified theoretical model.

5.3 Re-evaluating TD learning and RPE signals

In Chapter 2 of my dissertation, I showed how an RPE signal cannot explain the movement correlates found in DA responses of DMS-projecting populations. In Chapter 3, I argued that VTA DA populations may be signaling features related to the task, but all of the cue-period responses could be interpreted as RPE signals with respect to those features. Where does the classic TD learning model stand given our two results? What can or can't the RPE signal explain?

One way I sought to reconcile the two results was **Figure 3.8,** in which we expanded the "feature-specific" RPE model beyond reward PE, but PEs in general. We showed in **Figure 3.8d** that adapting a feature-specific RPE model to the task was unable to reproduce the contralateral preference we see in the data. We could not find functional heterogeneity because the Parker dataset used bulk photometry, and the signal we have would have been averaged over and thus could not show neuron-to-neuron specificity like we see in the Engelhard et al. dataset. Another reason could be that the reversal learning task is too simplistic; that is, given that the task did not require much complexity in the movement, there was no functional heterogeneity within the DMS-projecting DA population. This mirrors how in the original Schultz paper we only see a scalar RPE signal (Schultz, Apicella, and Ljungberg 1993), but in the virtual reality evidence accumulation task in the Engelhard paper we are able to see a richer heterogeneous DA signal (Engelhard et al. 2019).

In Chapter 4, I hypothesized that maybe DA signals PE in general, and that the movement correlates in the Parker dataset might not have been RPE with respect to movement, but value-free movement PEs instead. That is, while TD learning uses a reward PE, we cannot ignore the functional differences of projection targets, and thus we do need some modification of the TD learning model in order to address these movement correlates. I have some preliminary results in Chapter 4 that hint how such a movement signal might occur as an action surprise signal, but more analysis will need to be done to definitively show that it is indeed value-free and more theoretical work is needed to understand why and how such a signal would exist.

However, if the results from Chapter 4 do hold, I put forth the hypothesis that all DA might be signaling PE in general, but the PE may show up in different forms depending on the recording method, the task the animal engages in, and the projection targets of the recording sites. In the Engelhard dataset (Engelhard et al. 2019), we see functional heterogeneity with respect to reward because of the high-dimensional input state, the 2-photon neuron-resolution recording methods, and also because VTA DA neurons primarily signal reward. On the other hand, in the Parker dataset (Parker et al. 2016), we only find a scalar unsigned movement PE signal, given that the mice were learning a simpler task and that we used bulk-photometry to record dorsal-projecting DA neurons. Granted, this hypothesis that dopamine might reflect a universal prediction error signal has been proposed previously (Gardner, Schoenbaum, and Gershman 2018; Seitz et al. 2022). In my attempt, I try to preserve the core functionality of TD learning, but update how the PE can be formulated given the constraints of circuitry and projection-specific functionality.

Bibliography

Akiti, Korleki, Iku Tsutsui-Kimura, Yudi Xie, Alexander Mathis, Jeffrey E. Markowitz, Rockwell Anyoha, Sandeep Robert Datta, Mackenzie Weygandt Mathis, Naoshige Uchida, and Mitsuko Watabe-Uchida. 2022. "Striatal Dopamine Explains Novelty-Induced Behavioral Dynamics and Individual Variability in Threat Prediction." *Neuron* 110 (22): 3789–3804.e9.

Albin, R. L., A. B. Young, and J. B. Penney. 1989. "The Functional Anatomy of Basal Ganglia Disorders." *Trends in Neurosciences* 12 (10): 366–75.

Alexander, G. E., and M. D. Crutcher. 1990. "Functional Architecture of Basal Ganglia Circuits: Neural Substrates of Parallel Processing." *Trends in Neurosciences* 13 (7): 266–71.

Alexander, G. E., M. R. DeLong, and P. L. Strick. 1986. "Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex." *Annual Review of Neuroscience* 9: 357–81.

Anderegg, Angela, Jean-Francois Poulin, and Rajeshwar Awatramani. 2015. "Molecular Heterogeneity of Midbrain Dopaminergic Neurons--Moving toward Single Cell Resolution." *FEBS Letters* 589 (24 Pt A): 3714–26.

Arbuthnott, Gordon W., and Jeff Wickens. 2007. "Space, Time and Dopamine." *Trends in Neurosciences* 30 (2): 62–69.

Aronov, Dmitriy, and David W. Tank. 2014. "Engagement of Neural Circuits Underlying 2D Spatial Navigation in a Rodent Virtual Reality System." *Neuron* 84 (2): 442–56.

Avvisati, Riccardo, Anna-Kristin Kaufmann, Callum J. Young, Gabriella E. Portlock, Sophie Cancemi, Rui Ponte Costa, Peter J. Magill, and Paul D. Dodson. 2022. "Distributional Coding of Associative Learning within Projection-Defined Populations of Midbrain Dopamine Neurons." *bioRxiv*. https://doi.org/10.1101/2022.07.18.500429.

Baird, L. C. 1994. "Reinforcement Learning in Continuous Time: Advantage Updating." In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, 4:2448–53 vol.4.

Balleine, Bernard W., and John P. O'Doherty. 2010. "Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action." *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 35 (1): 48–69.

Barreto, André, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. 2016. "Successor Features for Transfer in Reinforcement Learning." *arXiv [cs.AI]*. arXiv. https://proceedings.neurips.cc/paper/2017/hash/350db081a661525235354dd3e19b8c05-Abstract.html.

Barter, Joseph W., Suellen Li, Dongye Lu, Ryan A. Bartholomew, Mark A. Rossi, Charles T. Shoemaker, Daniel Salas-Meza, Erin Gaidis, and Henry H. Yin. 2015. "Beyond Reward Prediction Errors: The Role of Dopamine in Movement Kinematics." *Frontiers in Integrative Neuroscience* 9 (May): 39.

Barto, A. G., R. S. Sutton, and C. W. Anderson. 1983. "Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13 (5): 834–46.

Barto, Andrew G. 1995. "1' 1 Adaptive Critics and the Basal Ganglia,.'" *Models of Information Processing in the Basal Ganglia*, 215.

Bayer, Hannah M., and Paul W. Glimcher. 2005. "Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal." *Neuron* 47 (1): 129–41.

Beckstead, R. M., V. B. Domesick, and W. J. Nauta. 1979. "Efferent Connections of the Substantia Nigra and Ventral Tegmental Area in the Rat." *Brain Research* 175 (2): 191–217.

Beier, Kevin T., Elizabeth E. Steinberg, Katherine E. DeLoach, Stanley Xie, Kazunari Miyamichi, Lindsay Schwarz, Xiaojing J. Gao, Eric J. Kremer, Robert C. Malenka, and Liqun Luo. 2015. "Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping." *Cell* 162 (3): 622–34.

Bellman, Richard. 1957. "A Markovian Decision Process." *Journal of Mathematics and Mechanics* 6 (5): 679–84.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Berke, Joshua D. 2018. "What Does Dopamine Mean?" *Nature Neuroscience* 21 (6): 787–93.

Beyene, Abraham G., Kristen Delevich, Jackson Travis Del Bonis-O'Donnell, David J. Piekarski, Wan Chen Lin, A. Wren Thomas, Sarah J. Yang, et al. 2019. "Imaging Striatal Dopamine Release Using a Nongenetically Encoded near Infrared Fluorescent Catecholamine Nanosensor." *Science Advances* 5 (7): eaaw3108.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2014. "Julia: A Fresh Approach to Numerical Computing." *arXiv [cs.MS]*. arXiv. http://arxiv.org/abs/1411.1607.

Björklund, Anders, and Stephen B. Dunnett. 2007. "Dopamine Neuron Systems in the Brain: An Update." *Trends in Neurosciences* 30 (5): 194–202.

Bogacz, Rafal. 2020. "Dopamine Role in Learning and Action Inference." *eLife* 9 (July). https://doi.org/10.7554/eLife.53262.

Boorman, Erie D., Timothy E. J. Behrens, Mark W. Woolrich, and Matthew F. S. Rushworth. 2009. "How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action." *Neuron* 62 (5): 733–43.

Bornstein, Aaron M., and Nathaniel D. Daw. 2011. "Multiplicity of Control in the Basal Ganglia: Computational Roles of Striatal Subregions." *Current Opinion in Neurobiology* 21 (3): 374–80.

Botvinick, Matthew M., Yael Niv, and Andew G. Barto. 2009. "Hierarchically Organized Behavior and Its Neural Foundations: A Reinforcement Learning Perspective." *Cognition* 113 (3): 262–80.

Bradfield, Laura A., and Genevra Hart. 2020. "Rodent Medial and Lateral Orbitofrontal Cortices Represent Unique Components of Cognitive Maps of Task Space." *Neuroscience and Biobehavioral Reviews* 108 (January): 287–94.

Brockman, Greg, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. "OpenAI Gym." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1606.01540.

Brück, Anna, Sargo Aalto, Elina Nurmi, Tero Vahlberg, Jörgen Bergman, and Juha O. Rinne. 2006. "Striatal Subregional 6-[18F]fluoro-L-Dopa Uptake in Early Parkinson's Disease: A Two-Year Follow-up Study." *Movement Disorders: Official Journal of the Movement*

Disorder Society 21 (7): 958–63.

Cai, Lili X., Katherine Pizano, Gregory W. Gundersen, Cameron L. Hayes, Weston T. Fleming, Sebastian Holt, Julia M. Cox, and Ilana B. Witten. 2020. "Distinct Signals in Medial and Lateral VTA Dopamine Neurons Modulate Fear Extinction at Different Times." *eLife* 9 (June). https://doi.org/10.7554/eLife.54936.

Caillé, I., B. Dumartin, and B. Bloch. 1996. "Ultrastructural Localization of D1 Dopamine Receptor Immunoreactivity in Rat Striatonigral Neurons and Its Relation with Dopaminergic Innervation." *Brain Research* 730 (1-2): 17–31.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76 (1): 1–32.

Chang, Chun Yun, Matthew Gardner, Maria Gonzalez Di Tillio, and Geoffrey Schoenbaum. 2017. "Optogenetic Blockade of Dopamine Transients Prevents Learning Induced by Changes in Reward Features." *Current Biology*. https://doi.org/10.1016/j.cub.2017.09.049.

Choi, Jung Yoon, Hee Jae Jang, Sharon Ornelas, Weston T. Fleming, Daniel Fürth, Jennifer Au, Akhil Bandi, Esteban A. Engel, and Ilana B. Witten. 2020. "A Comparison of Dopaminergic and Cholinergic Populations Reveals Unique Contributions of VTA Dopamine Neurons to Short-Term Memory." *Cell Reports*. https://doi.org/10.1016/j.celrep.2020.108492.

Choi, Kyuhyun, Eugenio Piasini, Edgar Díaz-Hernández, Luigim Vargas Cifuentes, Nathan T. Henderson, Elizabeth N. Holly, Manivannan Subramaniyan, Charles R. Gerfen, and Marc V. Fuccillo. 2023. "Distributed Processing for Value-Based Choice by Prelimbic Circuits Targeting Anterior-Posterior Dorsal Striatal Subregions in Male Mice." *Nature Communications* 14 (1): 1920.

Coddington, Luke T., and Joshua T. Dudman. 2018. "The Timing of Action Determines Reward Prediction Signals in Identified Midbrain Dopamine Neurons." *Nature Neuroscience*, October. https://doi.org/10.1038/s41593-018-0245-7.

Cohen, Jeremiah Y., Sebastian Haesler, Linh Vong, Bradford B. Lowell, and Naoshige Uchida. 2012. "Neuron-Type-Specific Signals for Reward and Punishment in the Ventral Tegmental Area." *Nature* 482 (7383): 85–88.

Collins, Anne G. E., and Michael J. Frank. 2014. "Opponent Actor Learning (OpAL): Modeling Interactive Effects of Striatal Dopamine on Reinforcement Learning and Choice Incentive." *Psychological Review* 121 (3): 337–66.

Collins, Anne L., and Benjamin T. Saunders. 2020. "Heterogeneity in Striatal Dopamine Circuits: Form and Function in Dynamic Reward Seeking." *Journal of Neuroscience Research* 98 (6): 1046–69.

Costall, B., R. J. Naylor, and C. Pycock. 1976. "Non-Specific Supersensitivity of Striatal Dopamine Receptors after 6-Hydroxydopamine Lesion of the Nigrostriatal Pathway." *European Journal of Pharmacology* 35 (2): 276–83.

Courville, Aaron C., Nathaniel D. Daw, and David S. Touretzky. 2005. "Similarity and Discrimination in Classical Conditioning: A Latent Variable Account." *Advances in Neural Information Processing Systems* 17: 313–20.

Cox, Julia, and Ilana B. Witten. 2019. "Striatal Circuits for Reward Learning and Decision-Making." *Nature Reviews. Neuroscience*, June. https://doi.org/10.1038/s41583-019-0189-2.

Dabney, Will, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis,

Rémi Munos, and Matthew Botvinick. 2020. "A Distributional Code for Value in Dopamine-Based Reinforcement Learning." *Nature* 577 (7792): 671–75.

Daw, Nathaniel D., Aaron C. Courville, and David S. Touretzky. 2003. "Timing and Partial Observability in the Dopamine System." In *Advances in Neural Information Processing Systems 15*, edited by S. Becker, S. Thrun, and K. Obermayer, 99–106. MIT Press.

———. 2006. "Representation and Timing in Theories of the Dopamine System." *Neural Computation* 18 (7): 1637–77.

Daw, Nathaniel D., Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. 2011. "Model-Based Influences on Humans' Choices and Striatal Prediction Errors." *Neuron* 69 (6): 1204–15.

Daw, Nathaniel D., Sham Kakade, and Peter Dayan. 2002. "Opponent Interactions between Serotonin and Dopamine." *Neural Networks: The Official Journal of the International Neural Network Society* 15 (4-6): 603–16.

Daw, Nathaniel D., Yael Niv, and Peter Dayan. 2005. "Uncertainty-Based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control." *Nature Neuroscience* 8 (12): 1704–11.

Daw, Nathaniel D., John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. 2006. "Cortical Substrates for Exploratory Decisions in Humans." *Nature* 441 (7095): 876–79.

DeLong, M. R. 1990. "Primate Models of Movement Disorders of Basal Ganglia Origin." *Trends in Neurosciences* 13 (7): 281–85.

Dickinson, A. 1985. "Actions and Habits: The Development of Behavioural Autonomy." *Philosophical Transactions of the Royal Society of London* 308 (1135): 67–78.

Diuk, Carlos, Karin Tsai, Jonathan Wallis, Matthew Botvinick, and Yael Niv. 2013. "Hierarchical Learning Induces Two Simultaneous, but Separable, Prediction Errors in Human Basal Ganglia." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 33 (13): 5797–5805.

Dodson, Paul D., Jakob K. Dreyer, Katie A. Jennings, Emilie C. J. Syed, Richard Wade-Martins, Stephanie J. Cragg, J. Paul Bolam, and Peter J. Magill. 2016. "Representation of Spontaneous Movement by Dopaminergic Neurons Is Cell-Type Selective and Disrupted in Parkinsonism." *Proceedings of the National Academy of Sciences of the United States of America* 113 (15): E2180–88.

Doeller, Christian F., John A. King, and Neil Burgess. 2008. "Parallel Striatal and Hippocampal Systems for Landmarks and Boundaries in Spatial Memory." *Proceedings of the National Academy of Sciences of the United States of America* 105 (15): 5915–20.

Dolan, Ray J., and Peter Dayan. 2013. "Goals and Habits in the Brain." *Neuron* 80 (2): 312–25.

Doll, Bradley B., Dylan A. Simon, and Nathaniel D. Daw. 2012. "The Ubiquity of Model-Based Reinforcement Learning." *Current Opinion in Neurobiology* 22 (6): 1075–81.

Doya, Kenji, Kazuyuki Samejima, Ken-Ichi Katagiri, and Mitsuo Kawato. 2002. "Multiple Model-Based Reinforcement Learning." *Neural Computation* 14 (6): 1347–69.

Engelhard, Ben, Joel Finkelstein, Julia Cox, Weston Fleming, Hee Jae Jang, Sharon Ornelas, Sue Ann Koay, et al. 2019. "Specialized Coding of Sensory, Motor and Cognitive Variables in VTA Dopamine Neurons." *Nature* 570 (7762): 509–13.

Eshel, Neir, Michael Bukwich, Vinod Rao, Vivian Hemmelder, Ju Tian, and Naoshige Uchida. 2015. "Arithmetic and Local Circuitry Underlying Dopamine Prediction Errors." *Nature*.

https://doi.org/10.1038/nature14855.
Eshel, Neir, Ju Tian, Michael Bukwich, and Naoshige Uchida. 2016. "Dopamine Neurons Share Common Response Function for Reward Prediction Error." *Nature Neuroscience* 19 (3): 479–86.
Estes, William K. 1950. "Toward a Statistical Theory of Learning." *Psychological Review* 57 (2): 94.
Frank, Michael J. 2005. "Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cognitive Deficits in Medicated and Nonmedicated Parkinsonism." *Journal of Cognitive Neuroscience* 17 (1): 51–72.
Frank, Michael J., and David Badre. 2012. "Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis." *Cerebral Cortex* 22 (3): 509–26.
Frank, Michael J., Lauren C. Seeberger, and Randall C. O'reilly. 2004. "By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism." *Science* 306 (5703): 1940–43.
Friston, K. J., G. Tononi, G. N. Reeke Jr, O. Sporns, and G. M. Edelman. 1994. "Value-Dependent Selection in the Brain: Simulation in a Synthetic Neural Model." *Neuroscience* 59 (2): 229–43.
Gardner, Matthew P. H., Geoffrey Schoenbaum, and Samuel J. Gershman. 2018. "Rethinking Dopamine as Generalized Prediction Error." *Proceedings. Biological Sciences / The Royal Society* 285 (1891). https://doi.org/10.1098/rspb.2018.1645.
Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 7 (4): 457–72.
Gershman, Samuel J. 2014. "Dopamine Ramps Are a Consequence of Reward Prediction Errors." *Neural Computation* 26 (3): 467–71.
Gershman, Samuel J., David M. Blei, and Yael Niv. 2010. "Context, Learning, and Extinction." *Psychological Review* 117 (1): 197–209.
Gershman, Samuel J., Kenneth A. Norman, and Yael Niv. 2015. "Discovering Latent Causes in Reinforcement Learning." *Current Opinion in Behavioral Sciences* 5 (October): 43–50.
Gershman, Samuel J., Bijan Pesaran, and Nathaniel D. Daw. 2009. "Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (43): 13524–31.
Gershman, Samuel J., and Geoffrey Schoenbaum. 2017. "Rethinking Dopamine Prediction Errors." *bioRxiv*. https://doi.org/10.1101/239731.
Gonon, F., J. B. Burie, M. Jaber, M. Benoit-Marand, B. Dumartin, and B. Bloch. 2000. "Geometry and Kinetics of Dopaminergic Transmission in the Rat Striatum and in Mice Lacking the Dopamine Transporter." *Progress in Brain Research* 125: 291–302.
Graybiel, Ann M. 2008. "Habits, Rituals, and the Evaluative Brain." *Annual Review of Neuroscience* 31 (1): 359–87.
Greenstreet, Francesca, Hernando Martinez Vergara, Sthitapranjya Pati, Laura Schwarz, Matthew Wisdom, Fred Marbach, Yvonne Johansson, et al. 2022. "Action Prediction Error: A Value-Free Dopaminergic Teaching Signal That Drives Stable Learning." *bioRxiv*. https://doi.org/10.1101/2022.09.12.507572.
Gremel, Christina M., and Rui M. Costa. 2013. "Orbitofrontal and Striatal Circuits Dynamically Encode the Shift between Goal-Directed and Habitual Actions." *Nature Communications* 4:

2264.

Gustafson, Nicholas J., and Nathaniel D. Daw. 2011. "Grid Cells, Place Cells, and Geodesic Generalization for Spatial Reinforcement Learning." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002235.

Guthrie, E. R. 1935. *Psychology of Learning* . Oxford, England: Harper.

Haber, S. N., J. L. Fudge, and N. R. McFarland. 2000. "Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 20 (6): 2369–82.

Hamid, Arif A., Michael J. Frank, and Christopher I. Moore. 2021. "Wave-like Dopamine Dynamics as a Mechanism for Spatiotemporal Credit Assignment." *Cell* 184 (10): 2733–49.e16.

Hart, Andrew S., Robb B. Rutledge, Paul W. Glimcher, and Paul E. M. Phillips. 2014. "Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (3): 698–704.

Hassan, Anhar, and Eduardo E. Benarroch. 2015. "Heterogeneity of the Midbrain Dopamine System." *Neurology*. https://doi.org/10.1212/wnl.0000000000002137.

Hill, Ashley, Antonin Raffin, Maximilian Ernestus, ↩adam Gleave, Anssi Kanervisto, Rene Traore, Prafulla Dhariwal, et al. 2018. "Stable Baselines." *GitHub Repository*. GitHub. https://github.com/hill-a/stable-baselines.

Hintiryan, Houri, Nicholas N. Foster, Ian Bowman, Maxwell Bay, Monica Y. Song, Lin Gou, Seita Yamashita, et al. 2016. "The Mouse Cortico-Striatal Projectome." *Nature Neuroscience* 19 (8): 1100–1114.

Hollerman, J. R., and W. Schultz. 1998. "Dopamine Neurons Report an Error in the Temporal Prediction of Reward during Learning." *Nature Neuroscience* 1 (4): 304–9.

Horvitz, J. C. 2000. "Mesolimbocortical and Nigrostriatal Dopamine Responses to Salient Non-Reward Events." *Neuroscience* 96 (4): 651–56.

Houk, J. C., J. L. Adams, and A. G. Barto. 1995. "A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement, Models of Information Processing in the Basal Ganglia (eds. JC Houk, JL Davis and DG Beiser), 249/270." MIT Press.

Howe, Mark W., and D. A. Dombeck. 2016. "Rapid Signalling in Distinct Dopaminergic Axons during Locomotion and Reward." *Nature* 535 (7613): 505–10.

Hunnicutt, Barbara J., Bart C. Jongbloets, William T. Birdsong, Katrina J. Gertz, Haining Zhong, and Tianyi Mao. 2016. "A Comprehensive Excitatory Input Map of the Striatum Reveals Novel Functional Organization." *eLife* 5 (November). https://doi.org/10.7554/eLife.19103.

Jiang, Linnie, and Ashok Litwin-Kumar. 2021. "Models of Heterogeneous Dopamine Signaling in an Insect Learning and Memory Center." *PLoS Computational Biology* 17 (8): e1009205.

Jin, Xin, and Rui M. Costa. 2010. "Start/stop Signals Emerge in Nigrostriatal Circuits during Sequence Learning." *Nature* 466 (7305): 457–62.

Jog, M. S., Y. Kubota, C. I. Connolly, V. Hillegaart, and A. M. Graybiel. 1999. "Building Neural Representations of Habits." *Science* 286 (5445): 1745–49.

Jones, Joshua L., Guillem R. Esber, Michael A. McDannald, Aaron J. Gruber, Alex Hernandez, Aaron Mirenzi, and Geoffrey Schoenbaum. 2012. "Orbitofrontal Cortex Supports Behavior

and Learning Using Inferred but Not Cached Values." *Science* 338 (6109): 953–56.

Jong, Johannes W. de, Seyedeh Atiyeh Afjei, Iskra Pollak Dorocic, James R. Peck, Christine Liu, Christina K. Kim, Lin Tian, Karl Deisseroth, and Stephan Lammel. 2019. "A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System." *Neuron* 101 (1): 133–51.e7.

Kalman, R. E. 1960. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering* 82 (1): 35–45.

Kamin, Leon J. 1967. "Attention-like Processes in Classical Conditioning." In *SYMP. ON AVERSIVE MOTIVATION MIAMI*. https://ntrs.nasa.gov/api/citations/19680013711/downloads/19680013711.pdf.

Kim, Hoseok, Jung Hoon Sul, Namjung Huh, Daeyeol Lee, and Min Whan Jung. 2009. "Role of Striatum in Updating Values of Chosen Actions." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (47): 14701–12.

Kim, Hyunggoo R., Athar N. Malik, John G. Mikhael, Pol Bech, Iku Tsutsui-Kimura, Fangmiao Sun, Yajun Zhang, et al. 2020. "A Unified Framework for Dopamine Signals across Timescales." *Cell* 183 (6): 1600–1616.e25.

Kishida, Kenneth T., Ignacio Saez, Terry Lohrenz, Mark R. Witcher, Adrian W. Laxton, Stephen B. Tatter, Jason P. White, Thomas L. Ellis, Paul E. M. Phillips, and P. Read Montague. 2016. "Subsecond Dopamine Fluctuations in Human Striatum Encode Superposed Error Signals about Actual and Counterfactual Reward." *Proceedings of the National Academy of Sciences of the United States of America* 113 (1): 200–205.

Kish, S. J., K. Shannak, and O. Hornykiewicz. 1988. "Uneven Pattern of Dopamine Loss in the Striatum of Patients with Idiopathic Parkinson's Disease. Pathophysiologic and Clinical Implications." *The New England Journal of Medicine* 318 (14): 876–80.

Kosillo, Polina, Yan-Feng Zhang, Sarah Threlfell, and Stephanie J. Cragg. 2016. "Cortical Control of Striatal Dopamine Transmission via Striatal Cholinergic Interneurons." *Cerebral Cortex* , August. https://doi.org/10.1093/cercor/bhw252.

Kremer, Yves, Jérôme Flakowski, Clément Rohner, and Christian Lüscher. 2020. "Context-Dependent Multiplexing by Individual VTA Dopamine Neurons." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, August. https://doi.org/10.1523/JNEUROSCI.0502-20.2020.

Kurth-Nelson, Zeb, and A. David Redish. 2009. "Temporal-Difference Reinforcement Learning with Distributed Representations." *PloS One* 4 (10): e7362.

Lak, Armin, Kensaku Nomoto, Mehdi Keramati, Masamichi Sakagami, and Adam Kepecs. 2017. "Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision." *Current Biology: CB* 27 (6): 821–32.

Lammel, Stephan, Daniela I. Ion, Jochen Roeper, and Robert C. Malenka. 2011. "Projection-Specific Modulation of Dopamine Neuron Synapses by Aversive and Rewarding Stimuli." *Neuron* 70 (5): 855–62.

Lammel, Stephan, Byung Kook Lim, Chen Ran, Kee Wui Huang, Michael J. Betley, Kay M. Tye, Karl Deisseroth, and Robert C. Malenka. 2012. "Input-Specific Control of Reward and Aversion in the Ventral Tegmental Area." *Nature* 491 (7423): 212–17.

Lau, Brian, and Paul W. Glimcher. 2008. "Value Representations in the Primate Striatum during Matching Behavior." *Neuron* 58 (3): 451–63.

Lau, Brian, Tiago Monteiro, and Joseph J. Paton. 2017. "The Many Worlds Hypothesis of

Dopamine Prediction Error: Implications of a Parallel Circuit Architecture in the Basal Ganglia." *Current Opinion in Neurobiology* 46 (October): 241–47.

Lee, Daeyeol, Hyojung Seo, and Min Whan Jung. 2012. "Neural Basis of Reinforcement Learning and Decision Making." *Annual Review of Neuroscience* 35 (March): 287–308.

Lee, Rachel S., Marcelo G. Mattar, Nathan F. Parker, Ilana B. Witten, and Nathaniel D. Daw. 2019. "Reward Prediction Error Does Not Explain Movement Selectivity in DMS-Projecting Dopamine Neurons." *eLife* 8 (April). https://doi.org/10.7554/eLife.42992.

Lee, Rachel S., Yotam Sagiv, Ben Engelhard, Ilana B. Witten, and Nathaniel D. Daw. 2023. "A Feature-Specific Prediction Error Model Explains Dopaminergic Heterogeneity." *bioRxiv*. https://doi.org/10.1101/2022.02.28.482379.

Lerner, Talia N., Carrie Shilyansky, Thomas J. Davidson, Kathryn E. Evans, Kevin T. Beier, Kelly A. Zalocusky, Ailey K. Crow, et al. 2015. "Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits." *Cell* 162 (3): 635–47.

Li, Jian, and Nathaniel D. Daw. 2011. "Signals in Human Striatum Are Appropriate for Policy Update rather than Value Prediction." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 31 (14): 5504–11.

Lindsey, Jack, and Ashok Litwin-Kumar. 2022. "Action-Modulated Midbrain Dopamine Activity Arises from Distributed Control Policies." *arXiv [q-bio.NC]*. arXiv. http://arxiv.org/abs/2207.00636.

Liu, Changliang, and Pascal S. Kaeser. 2019. "Mechanisms and Regulation of Dopamine Release." *Current Opinion in Neurobiology* 57 (August): 46–53.

Liu, Changliang, Lauren Kershberg, Jiexin Wang, Shirin Schneeberger, and Pascal S. Kaeser. 2018. "Dopamine Secretion Is Mediated by Sparse Active Zone-like Release Sites." *Cell* 172 (4): 706–18.e15.

Lloyd, Kevin, and Peter Dayan. 2016. "Safety out of Control: Dopamine and Defence." *Behavioral and Brain Functions: BBF* 12 (1): 15.

Marinelli, M., and J. E. McCutcheon. 2014. "Heterogeneity of Dopamine Neuron Activity across Traits and States." *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2014.07.034.

Matsuda, Wakoto, Takahiro Furuta, Kouichi C. Nakamura, Hiroyuki Hioki, Fumino Fujiyama, Ryohachi Arai, and Takeshi Kaneko. 2009. "Single Nigrostriatal Dopaminergic Neurons Form Widely Spread and Highly Dense Axonal Arborizations in the Neostriatum." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (2): 444–53.

Matsumoto, Hideyuki, Ju Tian, Naoshige Uchida, and Mitsuko Watabe-Uchida. 2016. "Midbrain Dopamine Neurons Signal Aversion in a Reward-Context-Dependent Manner." *eLife* 5 (October). https://doi.org/10.7554/eLife.17328.

Matsumoto, Masayuki, and Okihide Hikosaka. 2009. "Two Types of Dopamine Neuron Distinctly Convey Positive and Negative Motivational Signals." *Nature* 459 (7248): 837–41.

Matsumoto, N., T. Minamimoto, A. M. Graybiel, and M. Kimura. 2001. "Neurons in the Thalamic CM-Pf Complex Supply Striatal Neurons with Information about Behaviorally Significant Sensory Events." *Journal of Neurophysiology* 85 (2): 960–76.

McClure, Samuel M., Nathaniel D. Daw, and P. Read Montague. 2003. "A Computational Substrate for Incentive Salience." *Trends in Neurosciences* 26 (8): 423–28.

Menegas, William, Korleki Akiti, Ryunosuke Amo, Naoshige Uchida, and Mitsuko Watabe-Uchida. 2018. "Dopamine Neurons Projecting to the Posterior Striatum Reinforce

Avoidance of Threatening Stimuli." *Nature Neuroscience* 21 (10): 1421–30.

Menegas, William, Benedicte M. Babayan, Naoshige Uchida, and Mitsuko Watabe-Uchida. 2017. "Opposite Initialization to Novel Cues in Dopamine Signaling in Ventral and Posterior Striatum in Mice." *eLife* 6 (January). https://doi.org/10.7554/eLife.21886.

Mikhael, John G., Hyunggoo R. Kim, Naoshige Uchida, and Samuel J. Gershman. 2022. "The Role of State Uncertainty in the Dynamics of Dopamine." *Current Biology: CB*, January. https://doi.org/10.1016/j.cub.2022.01.025.

Miller, Kevin J., Matthew M. Botvinick, and Carlos D. Brody. 2017. "Dorsal Hippocampus Contributes to Model-Based Planning." *Nature Neuroscience* 20 (9): 1269–76.

Miller, Kevin, Amitai Shenhav, and Elliot Ludvig. 2019. "Habits without Values." *Psychological Review*, January, 067603.

Mnih, Volodymyr, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. "Asynchronous Methods for Deep Reinforcement Learning." In *International Conference on Machine Learning*, 1928–37. jmlr.org.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518 (7540): 529–33.

Montague, P. R., P. Dayan, and T. J. Sejnowski. 1996. "A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 16 (5): 1936–47.

Morris, Genela, Alon Nevet, David Arkadir, Eilon Vaadia, and Hagai Bergman. 2006. "Midbrain Dopamine Neurons Encode Decisions for Future Action." *Nature Neuroscience* 9 (8): 1057–63.

Moss, Morgane M., Peter Zatka-Haas, Kenneth D. Harris, Matteo Carandini, and Armin Lak. 2021. "Dopamine Axons in Dorsal Striatum Encode Contralateral Visual Stimuli and Choices." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, July. https://doi.org/10.1523/JNEUROSCI.0490-21.2021.

Niv, Yael. 2019. "Learning Task-State Representations." *Nature Neuroscience* 22 (10): 1544–53.

Niv, Yael, Nathaniel D. Daw, Daphna Joel, and Peter Dayan. 2007. "Tonic Dopamine: Opportunity Costs and the Control of Response Vigor." *Psychopharmacology* 191 (3): 507–20.

O'Doherty, John P. 2014. "The Problem with Value." *Neuroscience and Biobehavioral Reviews* 43 (June): 259–68.

O'Doherty, John P., Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. 2004. "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning." *Science* 304 (5669): 452–54.

Palminteri, Stefano, Thomas Boraud, Gilles Lafargue, Bruno Dubois, and Mathias Pessiglione. 2009. "Brain Hemispheres Selectively Track the Expected Value of Contralateral Options." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (43): 13465–72.

Palminteri, Stefano, Valentin Wyart, and Etienne Koechlin. 2017. "The Importance of Falsification in Computational Cognitive Modeling." *Trends in Cognitive Sciences* 21 (6): 425–33.

Pan, Weixing X., Tianyi Mao, and Joshua T. Dudman. 2010. "Inputs to the Dorsal Striatum of

the Mouse Reflect the Parallel Circuit Architecture of the Forebrain." *Frontiers in Neuroanatomy* 4 (December): 147.

Parker, Nathan F., Avinash Baidya, Julia Cox, Laura M. Haetzel, Anna Zhukovskaya, Malavika Murugan, Ben Engelhard, Mark S. Goldman, and Ilana B. Witten. 2022. "Choice-Selective Sequences Dominate in Cortical Relative to Thalamic Inputs to NAc to Support Reinforcement Learning." *Cell Reports* 39 (7): 110756.

Parker, Nathan F., Courtney M. Cameron, Joshua P. Taliaferro, Junuk Lee, Jung Yoon Choi, Thomas J. Davidson, Nathaniel D. Daw, and Ilana B. Witten. 2016. "Reward and Choice Encoding in Terminals of Midbrain Dopamine Neurons Depends on Striatal Target." *Nature Neuroscience* 19 (6): 845–54.

Parkinson, J. A., J. W. Dalley, R. N. Cardinal, A. Bamford, B. Fehnert, G. Lachenal, N. Rudarakanchana, K. M. Halkerston, T. W. Robbins, and B. J. Everitt. 2002. "Nucleus Accumbens Dopamine Depletion Impairs Both Acquisition and Performance of Appetitive Pavlovian Approach Behaviour: Implications for Mesoaccumbens Dopamine Function." *Behavioural Brain Research* 137 (1): 149–63.

Pearce, J. M., and G. Hall. 1980. "A Model for Pavlovian Learning: Variations in the Effectiveness of Conditioned but Not of Unconditioned Stimuli." *Psychological Review* 87 (6): 532–52.

Pinto, Lucas, Sue A. Koay, Ben Engelhard, Alice M. Yoon, Ben Deverett, Stephan Y. Thiberge, Ilana B. Witten, David W. Tank, and Carlos D. Brody. 2018. "An Accumulation-of-Evidence Task Using Visual Pulses for Mice Navigating in Virtual Reality." *Frontiers in Behavioral Neuroscience* 12 (March): 36.

Poulin, Jean-Francois, Giuliana Caronia, Caitlyn Hofer, Qiaoling Cui, Brandon Helm, Charu Ramakrishnan, C. Savio Chan, Daniel A. Dombeck, Karl Deisseroth, and Rajeshwar Awatramani. 2018. "Mapping Projections of Molecularly Defined Dopamine Neuron Subtypes Using Intersectional Genetic Approaches." *Nature Neuroscience* 21 (9): 1260–71.

Rao, Rajesh P. N. 2010. "Decision Making under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes." *Frontiers in Computational Neuroscience* 4 (November): 146.

Rescorla, R. A. 1972. "A Theory of Pavlovian Conditioning : Variations in the Effectiveness of Reinforcement and Nonreinforcement." *Current Research and Theory*, 64–99.

Reynolds, John N. J., B. I. Hyland, and J. R. Wickens. 2001. "A Cellular Mechanism of Reward-Related Learning." *Nature* 413 (6851): 67–70.

Reynolds, John N. J., and Jeffery R. Wickens. 2002. "Dopamine-Dependent Plasticity of Corticostriatal Synapses." *Neural Networks: The Official Journal of the International Neural Network Society* 15 (4-6): 507–21.

Ribas-Fernandes, José J. F., Alec Solway, Carlos Diuk, Joseph T. McGuire, Andrew G. Barto, Yael Niv, and Matthew M. Botvinick. 2011. "A Neural Signature of Hierarchical Reinforcement Learning." *Neuron* 71 (2): 370–79.

Rice, Margaret E., and Jyoti C. Patel. 2015. "Somatodendritic Dopamine Release: Recent Mechanistic Insights." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1672). https://doi.org/10.1098/rstb.2014.0185.

Robbe, David. 2018. "To Move or to Sense? Incorporating Somatosensory Representation into Striatal Functions." *Current Opinion in Neurobiology* 52 (October): 123–30.

Roesch, Matthew R., Donna J. Calu, and Geoffrey Schoenbaum. 2007. "Dopamine Neurons

Encode the Better Option in Rats Deciding between Differently Delayed or Sized Rewards." *Nature Neuroscience* 10 (12): 1615–24.

Rowland, Mark, Robert Dadashi, Saurabh Kumar, Remi Munos, Marc G. Bellemare, and Will Dabney. 2019. "Statistics and Samples in Distributional Reinforcement Learning." In *Proceedings of the 36th International Conference on Machine Learning*, edited by Kamalika Chaudhuri and Ruslan Salakhutdinov, 97:5528–36. Proceedings of Machine Learning Research. PMLR.

Rueda-Orozco, Pavel E., and David Robbe. 2015. "The Striatum Multiplexes Contextual and Kinematic Information to Constrain Motor Habits Execution." *Nature Neuroscience* 18 (3): 453–60.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–36.

Russek, Evan M., Ida Momennejad, Matthew M. Botvinick, Samuel J. Gershman, and Nathaniel D. Daw. 2017. "Predictive Representations Can Link Model-Based Reinforcement Learning to Model-Free Mechanisms." *PLoS Computational Biology* 13 (9): e1005768.

Russell, Stuart, and Andrew L. Zimdars. 2003. "Q-Decomposition for Reinforcement Learning Agents." In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 656–63. ICML'03. Washington, DC, USA: AAAI Press.

Samejima, Kazuyuki, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. 2005. "Representation of Action-Specific Reward Values in the Striatum." *Science* 310 (5752): 1337–40.

Samuelson, P. A. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5 (17): 61–71.

Saunders, Benjamin T., Jocelyn M. Richard, Elyssa B. Margolis, and Patricia H. Janak. 2018. "Dopamine Neurons Create Pavlovian Conditioned Stimuli with Circuit-Defined Motivational Properties." *Nature Neuroscience* 21 (8): 1072–83.

Schultz, W. 1998. "Predictive Reward Signal of Dopamine Neurons." *Journal of Neurophysiology* 80 (1): 1–27.

Schultz, W., P. Apicella, and T. Ljungberg. 1993. "Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli during Successive Steps of Learning a Delayed Response Task." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 13 (3): 900–913.

Schultz, W., P. Dayan, and P. R. Montague. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593–99.

Seitz, Benjamin M., Ivy B. Hoang, Aaron P. Blaisdell, and Melissa J. Sharpe. 2022. "Learning in Reverse: Dopamine Errors Drive Excitatory and Inhibitory Components of Backward Conditioning in an Outcome-Specific Manner." *bioRxiv*. https://doi.org/10.1101/2022.01.10.475719.

Shohamy, Daphna. 2011. "Learning and Motivation in the Human Striatum." *Current Opinion in Neurobiology* 21 (3): 408–14.

Silva, Joaquim Alves da, Fatuel Tecuapetla, Vitor Paixão, and Rui M. Costa. 2018. "Dopamine Neuron Activity before Action Initiation Gates and Invigorates Future Movements." *Nature* 554 (7691): 244–48.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–89.

Singh, Satinder Pal. 1992. "Transfer of Learning by Composing Solutions of Elemental Sequential Tasks." *Machine Learning* 8 (3): 323–39.

Soares, Sofia, Bassam V. Atallah, and Joseph J. Paton. 2016. "Midbrain Dopamine Neurons Control Judgment of Time." *Science* 354 (6317): 1273–77.

Stachenfeld, Kimberly L., Matthew M. Botvinick, and Samuel J. Gershman. 2017. "The Hippocampus as a Predictive Map." *Nature Neuroscience* 20 (11): 1643–53.

Steinberg, Elizabeth E., Ronald Keiflin, Josiah R. Boivin, Ilana B. Witten, Karl Deisseroth, and Patricia H. Janak. 2013. "A Causal Link between Prediction Errors, Dopamine Neurons and Learning." *Nature Neuroscience* 16 (7): 966–73.

Suri, R. E., and W. Schultz. 1999. "A Neural Network Model with Dopamine-like Reinforcement Signal That Learns a Spatial Delayed Response Task." *Neuroscience* 91 (3): 871–90.

Sutton, Richard S. 1988. "Learning to Predict by the Methods of Temporal Differences." *Machine Learning* 3 (1): 9–44.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning, Second Edition: An Introduction*. MIT Press.

Sutton, Richard S., Andrew G. Barto, and Others. 1998. *Introduction to Reinforcement Learning*. Vol. 135. MIT press Cambridge.

Swanson, L. W. 1982. "The Projections of the Ventral Tegmental Area and Adjacent Regions: A Combined Fluorescent Retrograde Tracer and Immunofluorescence Study in the Rat." *Brain Research Bulletin* 9 (1-6): 321–53.

Syed, Emilie C. J., Laura L. Grima, Peter J. Magill, Rafal Bogacz, Peter Brown, and Mark E. Walton. 2016. "Action Initiation Shapes Mesolimbic Dopamine Encoding of Future Rewards." *Nature Neuroscience* 19 (1): 34–36.

Takahashi, Yuji K., Hannah M. Batchelor, Bing Liu, Akash Khanna, Marisela Morales, and Geoffrey Schoenbaum. 2017. "Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards." *Neuron* 95 (6): 1395–1405.e3.

Takahashi, Yuji K., Matthew R. Roesch, Robert C. Wilson, Kathy Toreson, Patricio O'Donnell, Yael Niv, and Geoffrey Schoenbaum. 2011. "Expectancy-Related Changes in Firing of Dopamine Neurons Depend on Orbitofrontal Cortex." *Nature Neuroscience* 14 (12): 1590–97.

Takahashi, Yuji K., and Geoffrey Schoenbaum. 2016. "Ventral Striatal Lesions Disrupt Dopamine Neuron Signaling of Differences in Cue Value Caused by Changes in Reward Timing but Not Number." *Behavioral Neuroscience* 130 (6): 593–99.

Takahashi, Yuji, Geoffrey Schoenbaum, and Yael Niv. 2008. "Silencing the Critics: Understanding the Effects of Cocaine Sensitization on Dorsolateral and Ventral Striatum in the Context of an Actor/critic Model." *Frontiers in Neuroscience* 2 (1): 86–99.

Tano, Pablo, Peter Dayan, and Alexandre Pouget. 2020. "A Local Temporal Difference Code for Distributional Reinforcement Learning." *Advances in Neural Information Processing Systems* 33.
https://proceedings.neurips.cc/paper/2020/file/9dd16e049becf4d5087c90a83fea403b-Paper.pdf.

Tricomi, Elizabeth, Bernard W. Balleine, and John P. O'Doherty. 2009. "A Specific Role for Posterior Dorsolateral Striatum in Human Habit Learning." *The European Journal of Neuroscience* 29 (11): 2225–32.

Tsujimoto, Satoshi, Aldo Genovesio, and Steven P. Wise. 2011. "Frontal Pole Cortex: Encoding

Ends at the End of the Endbrain." *Trends in Cognitive Sciences* 15 (4): 169–76.
Tsutsui-Kimura, Iku, Hideyuki Matsumoto, Naoshige Uchida, and Mitsuko Watabe-Uchida. 2020. "Distinct Temporal Difference Error Signals in Dopamine Axons in Three Regions of the Striatum in a Decision-Making Task." https://doi.org/10.1101/2020.08.22.262972.
Ungerstedt, U., and G. W. Arbuthnott. 1970. "Quantitative Recording of Rotational Behavior in Rats after 6-Hydroxy-Dopamine Lesions of the Nigrostriatal Dopamine System." *Brain Research* 24 (3): 485–93.
Ungless, Mark A., Peter J. Magill, and J. Paul Bolam. 2004. "Uniform Inhibition of Dopamine Neurons in the Ventral Tegmental Area by Aversive Stimuli." *Science* 303 (5666): 2040–42.
Verharen, J. P. H., Yichen Zhu, and Stephan Lammel. 2020. "Aversion Hot Spots in the Dopamine System." *Current Opinion in Neurobiology* 64 (October): 46–52.
Von Neumann, John, and Oskar Morgenstern. 1947. "Theory of Games and Economic Behavior, 2nd Rev." https://trafficlight.bitdefender.com/info?url=https://psycnet.apa.org/record/1947-03159-000.
Voorn, Pieter, Louk J. M. J. Vanderschuren, Henk J. Groenewegen, Trevor W. Robbins, and Cyriel M. A. Pennartz. 2004. "Putting a Spin on the Dorsal–ventral Divide of the Striatum." *Trends in Neurosciences* 27 (8): 468–74.
Wei, Wei, Ali Mohebi, and Joshua D. Berke. 2021. "Striatal Dopamine Pulses Follow a Temporal Discounting Spectrum." *bioRxiv*. https://doi.org/10.1101/2021.10.31.466705.
Widrow, Bernard, and Marcian E. Hoff. 1960. "Adaptive Switching Circuits." In *IRE WESCON Convention Record*, 4:96–104. New York.
Wilson, Robert C., Yuji K. Takahashi, G. Schoenbaum, and Yael Niv. 2014. "Orbitofrontal Cortex as a Cognitive Map of Task Space." *Neuron* 81 (2): 267–79.
Wise, Roy A. 2004. "Dopamine, Learning and Motivation." *Nature Reviews. Neuroscience* 5 (6): 483–94.
Wood, Wendy, and David T. Neal. 2007. "A New Look at Habits and the Habit-Goal Interface." *Psychological Review* 114 (4): 843–63.
Wood, Wendy, and Dennis Rünger. 2016. "Psychology of Habit." *Annual Review of Psychology* 67: 289–314.
Yin, Henry H., and Barbara J. Knowlton. 2006. "The Role of the Basal Ganglia in Habit Formation." *Nature Reviews. Neuroscience* 7 (6): 464–76.
Yin, Henry H., Barbara J. Knowlton, and Bernard W. Balleine. 2004. "Lesions of Dorsolateral Striatum Preserve Outcome Expectancy but Disrupt Habit Formation in Instrumental Learning." *The European Journal of Neuroscience* 19 (1): 181–89.
Yin, Henry H., Sean B. Ostlund, Barbara J. Knowlton, and Bernard W. Balleine. 2005. "The Role of the Dorsomedial Striatum in Instrumental Conditioning." *The European Journal of Neuroscience* 22 (2): 513–23.
Yttri, Eric A., and Joshua T. Dudman. 2016. "Opponent and Bidirectional Control of Movement Velocity in the Basal Ganglia." *Nature* 533 (7603): 402–6.
Zolin, Aryeh, Raphael Cohn, Rich Pang, Andrew F. Siliciano, Adrienne L. Fairhall, and Vanessa Ruta. 2021. "Context-Dependent Representations of Movement in Drosophila Dopaminergic Reinforcement Pathways." *Nature Neuroscience* 24 (11): 1555–66.