Dopamine Contributions to Motivational Vigor and Reinforcement Driven Learning

by

Arif Hamid

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Neuroscience)
in the University of Michigan
2016

Doctoral Committee:

Professor Joshua D. Berke, Chair
Professor Kent C. Berridge
Associate Professor Anatol Kreitzer, University of California, San Francisco
Assistant Professor Daniel K. Leventhal
Professor Satinder Singh

DEDICATION

For Babisha, Hoyii, Shafiye, MardoshQ, Linu, Ayu and Ebsuye;
Lubuu Tiya, Onee Tiyaa.
My heroes, my rocks to lean on.

ACKNOWLEDGMENTS

I am deeply indebted to many individuals who have been tremendously generous with their support, guidance, time, patience, rigor and friendship during (and before) my time in Ann Arbor. I am most grateful for my mentor Dr. Joshua Berke, who first of all, gave me a *fair* opportunity to succeed. Josh was relentless in pushing me to realize my dreams, and was always my most supportive and most critical voice for the entirety of my dissertation. The long hours we spent in his office on countless occasions, debating and strategizing the best way to study how the brain works remain the most gratifying and treasured moments of my adult life. Thank you Josh! I am also grateful for Dr. Brandon Aragon's patience, continued mentorship, and actively increasing my visibility and network at conferences.

I would like to thank my thesis committee members (Dr. Kent Berridge, Dr. Satinder Singh, Dr. Anatol Kreitzer and Dr. Daniel Leventhal), who provided many rounds of input for various project-ideas and manuscripts. I would also like to acknowledge Dr. Terry Robinson and Dr. Martin Sarter for energized conversations about dopamine and the brain in general.

I have been very fortunate to overlap with several amazing scientists on the 4th floor of East Hall that have helped shape my PhD intellectually, and also made my work a rich personal experience. These include Dr. Nicolas Mallet, Dr. Robert Schmidt, Dr. Ali Mohebi, Dr. Shanna Resendez, Caitlin Vander Weele, Dr. Vedran Lovic, Dr. Jeffrey Pettibone, Dr. BonMi Gu, Dr. Bryan Singer, Dr. Kyle Pitches, Dr. Kirsten Porter-Stransky, Alex Kawa, Pavlo Popov, Sofia Lopez, and many others.

Support from the graduate program in Neuroscience at Michigan was magnificent, and I especially want to thank Rachel Flaten, Valerie Smith and Dr. Edward Stuenkel.

I would like to acknowledge the lasting impact of my past mentors that have helped me crystalize, and attain my dreams of being a scientist. Dr. Eugenia Paulus, Dr. Peggy LePage and Dr. Johanna Abrams sparked a curiosity for experimentation and fanned its flames at North Hennepin Community College. Dr. Janet Dubinsky, Dr. Eric Newman, Dr. A. David Redish, Dr. Mark Masino and Dr. Anusha Mishra at the University of Minnesota incubated my curiosity and ensured I realized my goals. To all, I am eternally grateful, and promise to pay it forward.

Most importantly, I want to acknowledge the love and support from my family that made all of this possible. My father Ahmed Hamid and mother, Hindiya Dawe have sacrificed so much, and moved mountains to provide my family with an opportunity to study in the US, and continue to believe in me every second. My siblings Shafi, Mawerdi, Lina, Ayantu and Ebesa have always formed a legion of my personal heroes that always motivate, inspire and entertain me; thank you, I love you, I can never repay you.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Brain mechanisms for reinforcement learning and adaptive decision-making are widely accepted to critically involve the basal ganglia (BG) and the neurotransmitter dopamine (DA). DA is a key modulator of synaptic plasticity within the striatum, critically regulating neurophysiological adaptations for normal reinforcement driven learning, and maladaptive changes during disease conditions (e.g. drug addiction, Parkinson's disease). Activity in midbrain DA cells are reported to encode errors in reward prediction, providing a learning signal to guide future behaviors. Yet, dopamine is also a key modulatory of motivation, invigorating current behavior. Prevailing theories of DA emphasize its role in either affecting current performance, or modulating reward-related learning.

This thesis will present data aimed at resolving gaps in the literature for how DA makes simultaneous contributions to dissociable learning and motivational processes. Specifically, I argue that striatal DA fluctuations signal a single decision variable: a Value function (an ongoing estimate of discounted future rewards) that is used for motivational decision making ('Is It worth it?') and that abrupt deflections in this value function serve as temporal-difference reward prediction errors used for reinforcement/learning ("repeat action?"). These DA prediction errors may be causally involved in strengthening some, but not all, valuation mechanisms. Furthermore, DA activity on the midbrain-forebrain axis indicate a dissociation between DA cell bodies and their striatal terminals. I propose that this is an adaptive computational strategy, whereby DA targets tailor release to their own computational requirements, potentially converting an RPE-like spike signal into a motivational (value) message.

CHAPTER 1: GENERAL INTRODUCTION

The Basal Ganglia (BG), a collection of subcortical nuclei, are key in generating and executing a range of goal-directed behaviors. It is now widely accepted that the likely overall function of the BG is to decide what is worth doing (action-selection or action-energization) based on what has worked before (reinforcement learning). Much of the early evidence in support of this view comes from neurological and neuropsychiatric disorders that impact the BG.

To gain further insight into brain-mechanisms for goal-directed behavior and decision-making, it is important to understand the underlying anatomy and physiology of associated circuits. With this in mind, I will briefly review the basic anatomy and models of BG function as it pertains to this thesis, but review here is not exhaustive.

Basal ganglia anatomy

The BG nuclei include the striatum (Str, or caudate and putamen in primates, CaPu), pallidal nuclei (GPi and GPe respectively for internal and external segments of globus pallidus, and VP for ventral pallidus), sub-thalamic nucleus (STN) and the substantial nigra pars reticulata (SNr) (Gerfen and Bolam, 2010; Nelson and Kreitzer, 2014) (Figure 1.1). The striatum is the major input nucleus of the BG, receiving projections from many cortical areas related to limbic, motor and sensory functions. On the other hand, the SNr and GPi form the major output nuclei of BG, and send projections to thalamic relay nuclei (GPi projects to ventrolateral (vl) and parafascicular(pf) nuclei, and the SNr projects to ventromedial (vm), parafascicular (pf) and mediodorsal nuclei). SNr also sends direct projections to midbrain motor centers, namely the superior colliculus (SC) and pedunclopontine nucleus (PPN).

BG macro-circuits are organized in at least two respects (Middleton, 2000; Alexander et al., 1986; Redgrave et al., 2010) (Figure 1.2). First, loops of the cortico-BG are broadly classified into repeating cognitive, limbic and sensorimotor modules (Voorn et al., 2004). These modules can be delineated mainly based on source of cortical input, and, it appears that they subserve hierarchically organized classes of adaptive behaviors (Graybiel, 1998).

Second, inter-BG connectivity forms three major pathways (i.e direct, indirect, and hyperdirect) for information flow. (i) Medium spiny neurons (MSNs) of the striatum send axons directly to the output nucleus, SNr, forming the *'direct pathway'* (ii) A different group of MSNs reach the BG output via several relay nuclei. First, MSNs project to the GPe, which in turn synapses onto the STN. STN cells finally send axons to the SNr, completing the circuit for the *'indirect pathway'*. (iii) The *'hyperdirect pathway'* is composed of cortical neurons that send direct projections to the STN, bypassing the striatum (Frank, 2006).

Midbrain dopamine (DA) neurons are also a key component of BG circuitry. While they do not directly participate in BG fast neurotransmission, they robustly modulate activity and plasticity within various brain areas, and are critical for sensorimotor functions. DA cells reside in the ventral tegmental area (VTA) and substantial nigra pars compacta (SNc) midbrain nuclei, and make dense projections to the striatum with highly arborized axons.

The 'Box-and-arrow' model of BG function

One of the earliest integrated models of basal BG was the 'box-and-arrow' model, (also known as 'rate model') suggested by the Roger Albin and Mahlon DeLong (Albin et al., 1989; DeLong, 1990; DeLong, 1983; Penney and Young, 1983). The Albin-DeLong model describes the basic connectivity of BG, and postulates that aberrant activity in BG dynamics produce clinical hallmarks of movement disorders in parkinson's disease (PD) patients. Details of this model are presented below, but, I must first note that the BG nuclei principally use the inhibitory neurotransmitter GABA, and function via

disinhibition. In fact, all major nuclei make inhibitory projections, with the exception of the STN, which makes excitatory glutamate connections (Nelson and Kreitzer, 2014).

The *direct pathway* generally facilitates movement and is synonymously called the 'GO' pathway. Specifically, excitatory drive from the overlying cortex activates the inhibitory direct-pathway striatal MSNs (dMSNs) to mono-synaptically suppress activity in the SNr (Alexander et al., 1986; Freeze et al., 2013). SNr units tonically inhibit thalamo-cortical projections, and their transient suppression disinhibits the thalamus to promote movement (Hikosaka et al., 2000).

Conversely, the *indirect-pathway* is known as the 'NO-GO' pathway, and inhibits movement. Indirect-pathway MSNs (iMSNs) are stimulated by cortical glutamergic input, and release GABA to inhibit the GPe (Kita, 2007). Pauses in GPe activity disinhibits the STN, which in turn excites the SNr, resulting in the suppression of activity in the thalamus. Thus, over-activity of the BG output via SNr are thought to produce bradykinesia, akinesia and freezing in PD.

The direct/indirect dichotomy was initially supported by studies that identified markers enriched in cells of the two pathways. For example, dMSNs express dynorphin and substance-P in high levels (Beckstead and Kersey, 1985; Haber and Watson, 1983; Gerfen and Young, 1988), but, iMSNs are enriched with enkephalin (Penny et al., 1986; Gerfen and Young, 1988; Berendse et al., 1992b; Berendse et al., 1992a). Another cellular marker to distinguish the two striatal pathways is the localization of mRNA coding for two classes of DA receptors (Gerfen et al., 1990; Surmeier et al., 1996). dMSNs selectively express the D1-like DA receptors and iMSNs express D2-like receptors. The GO and NOGO pathways are now synonymously referred as the D1 and D2 pathways respectively.

Several predictions of the 'box-and-arrow' model were confirmed in electrophysiological characterization of BG nuclei during limb movement (Mink, 1996), generation of saccadic eye movements (Hikosaka et al., 2000), and temporally precise optogenetic studies (Freeze et al., 2013; Kravitz et al., 2010; Roseberry et al., 2016;

Tecuapetla et al., 2016). Specifically, direct optical excitation of genetically defined D1MSNs promote movement initiation, whereas stimulation of D2 MSNs decrease locomotion (Kravitz et al., 2010; Yttri and Dudman, 2016; Tecuapetla et al., 2016). In addition, excitation of dMSNs or iMSNs led to the (respective) inhibition and excitation of SNr spiking (Freeze et al., 2013). Together with the box-and-arrow framework, these studies further our understanding of BG circuit dynamics that underlie normal action generation and execution.

Despite strong working-frameworks, the precise temporal dynamics of the BG during action-selection/generation/execution, and a unified account of their role in movement continues to be refined. For instance, unlike previous suggestions that direct and indirect pathways compete for movement control, some recent reports suggest that the two pathways are concurrently active and together, modulate activity of BG output nuclei (Cui et al., 2013; Tecuapetla et al., 2016).

The ‘box-and-arrow’ view has been extended to encompass valuation mechanisms that influence selection of motor programs among many candidates (Dayan and Niv, 2008; Niv, 2009; O'Reilly and Frank, 2006). Action patterns that produce desirable results are more likely to be repeated when that situation is re-experienced. In this view, the computational roles of BG direct and indirect pathways are to arbitrate among candidate action-values and select the appropriate responses (Maia and Frank, 2011). In the brain, action-values are hypothesized to be instantiated in the synaptic strength of cortical inputs onto both iMSNs and dMSNs. Asymmetrical cortical drive of the GO and NO-GO activity are thought to dictate the likelihood of gating and selection of cortical motor programs based on striatal cached-values for those actions (Frank, 2011). Indeed, this view is supported by electrophysiological reports of action-value coding in striatal MSNs (Samejima et al., 2005). Furthermore, optogenetic stimulation of D1MSNs promote (and D2MSN stimulation discourages) choice in a manner consistent with increases (or decreases) in stored action-values (Tai et al., 2012).

Dynamic plasticity of glutamergic synapses onto GO and NOGO pathways hypothesized to underlie value modifications to affect future arbitration. It is now widely accepted that BG plasticity mechanisms (especially those involving rewards) are under strong control of striatal DA levels.

DA modulation of plasticity in BG decision circuit

DA neurons clustered in three midbrain groups send long-range axons along the medial forebrain bundle to modulate activity in many different brain regions (Dahlström and Fuxe, 1964; Lindvall et al., 1974; Fallon, 1988; Björklund and Lindvall, 1984; Gerfen et al., 1987; Haber et al., 2000). The A9 cluster of the SNc projects to the medial/lateral dorsal striatum (also known as nigrostriatal DA system), and the A8 and A10 groups of midbrain DA neurons project to the ventral striatum and frontal cortical regions (known as mesolimbic and mesocortical pathways respectively). Within the striatum, DA axons arborize extensively (Matsuda et al., 2009) to form a synapse every 1-2 μm^3, making up ~10% of all striatal synapses. (Descarries et al., 1996; Ingham et al., 1998).

Functionally, DA cells display two primary modes of firing pattens (Grace, 1991; Grace et al., 2007). A tonic, irregular discharge rate of 1-10 Hz in the absence of synaptic input appears to be driven by the combination of voltage gated calcium channels and calcium-gated potassium currents (Grace and Bunney, 1984b; Amini et al., 1999; Grace, 1991). On the other hand, stimulus-linked synaptic inputs synchronize DA cells and drive phasic discharge of a few action potentials at 15-100 Hz (Grace and Bunney, 1984a; Schultz, 2007).

Synaptic release of DA is predominantly produced by quantal fusion of synaptic vesicles at axonal boutons following spiking events (Sulzer et al., 2010; Sulzer, 2011). While excitation-secretion in DA cells are extensively regulated (e.g. via local micro-circuitry (Rice and Cragg, 2004; Threlfell et al., 2012; Cachope et al., 2012) and intracellular mechanisms (Pereira et al., 2016)), tonic and phasic spiking are hypothesized to influence either the slow-changing ambient levels or brief efflux of DA into the synaptic cleft. '*Tonic*' striatal DA levels are estimated to be 10-20 nM (Shou et

al., 2006) and are likely a byproduct of release-reuptake equilibria (Floresco et al., 2003). Synchronous bursts of DA cells briefly increases striatal levels by an additional 50-100nM, that are regulated in their temporal and spatial spread (Aragona et al., 2009; Wightman et al., 2007).

Postsynaptic effects of DA are mediated by G-protein coupled receptors broadly classified into D1-like and D2-like receptors (Gerfen and Surmeier, 2011; Kreitzer, 2009). DA produces immediate (and transient) modifications of MSN biophysical dynamics, in addition to persistent reorganization of synaptic inputs mediated by the two DA receptor classes. Specifically, D1 receptors have low-affinity for DA (Richfield et al., 1989), and increase excitability of striatal dMSNs (Surmeier et al., 1995; Kitai and Surmeier, 1992; Gao et al., 1997; Galarraga et al., 1997). On the other hand, D2 receptors have a high-affinity for DA (Richfield et al., 1989) and mediate decreases in cellular excitability via closure of L-type calcium channels and attenuation of AMPA currents (Hernández-López et al., 2000; Hernández-Echeagaray et al., 2004; Cepeda et al., 1993; Higley and Sabatini, 2010; Day et al., 2008).

D1 and D2 receptors are also responsible for long term potentiation (LTP) and long term depression (LTD) of synapses via complicated intracellular cascades. Briefly, potentiation of glutamergic synapses onto dMSNs are dependent on the D1 receptor (Shen et al., 2008; Pawlak and Kerr, 2008), which recruit adenyl-cyclase, PKA and MAP kinases (Kebabian and Greengard, 1971; Centonze et al., 2003). These intracellular cascades promote a variety of processes that stabilize and strengthen synapses to promote dMSN LTP (Calabresi et al., 2007; Kerr and Wickens, 2001; Shen et al., 2008; Yagishita et al., 2014), including insertion of ion channels, synthesis of new receptors and cytoskeletal stabilization of dendritic spines (Sweatt, 2004). D2 mediated LTD of glutamegic inputs are initiated postsynaptically, but primarily produce changes in the efficacy of presynaptically release. D2-LTD involves the co-activation of metabotropic glutamate receptors (mGluRs) and voltage gated calcium channels (VGCCs) to recruit endocanabiniods (CB), which then diffuse to presynaptic CB1 receptors to blunt glutamate release (Wang et al., 2006; Surmeier et al., 1995; Kreitzer and Malenka,

2007; Shen et al., 2008). Together these opposing intracellular cascades initiated by the two receptor types produce diverging direct/indirect circuit dynamics in response to DA neuromodulation.

If BG dynamics are exquisitely sensitive to DA release, and DA can produce profound synaptic rearrangement within the striatum, how are DA cells recruited? What environmental and behavioral conditions promote DA release? Furthermore, what are the behavioral consequences of DA release, and what specific brain mechanisms underlie these changes in behavior?

Role of DA in reward, motivation, and learning

These questions have been explored across many disciplines for several decades, and remain a subject of intense debate. A consensus account for DA function likely includes the processing of reward-related information, mediating reward-learning and modulation of reward-seeking behaviors (Berridge, 2007; Wise, 2004; Schultz, 2007; Dayan, 2009; Redish, 2004). However, the precise contribution how DA apparently modulates all of these processes remains a puzzle.

It is extensively documented that midbrain DA neurons are strongly recruited (and DA released into target regions) in response to primary rewards. For example, several studies have reported elevation of extracellular DA in animals receiving food (Roitman et al., 2004b), engaged in sex (Fiorino et al., 1997; Robinson et al., 2001), administration of drugs of abuse (Kiyatkin et al., 1993; Phillips et al., 2003) or even exploring novel environments (Rebec et al., 1997). It is now clear that DA is not involved in the hedonic aspect of consuming rewards (Wise, 1982), but is rather critical for *forming associations* about reward and their predictive cues. In addition DA appears to be critical for the *expression* of (immediate and future) motivated reward-seeking behaviors (Berridge and Robinson, 1998; Berridge and Robinson, 2003; Salamone and Correa, 2012; Redgrave et al., 2008; Schultz, 2015).

Interfering with brain DA signaling strongly affects immediate (online) motivation. For example intravenous delivery of cocaine or amphetamine (which cause DA efflux) produce robust psychomotor activation, heightened motivational arousal and alertness (Wise and Bozarth, 1987; Ahmed and Cador, 2006). Further, direct intracranial excitation of DA supports self-stimulation, wherein rats fervently respond to receive stimulation (Rolls et al., 1980; Wise, 1996; Witten et al., 2011). Pharmacological inhibition of DA function (via manipulation of DA transport, re-uptake or receptor occupancy) blunts the vigorous exertion of effort to pursue and obtain rewards (Salamone et al., 2009). Finally, DA depleted mice display profound deficits in movement and motivation to seek rewards (Zhou and Palmiter, 1995), suggesting a critical DA contribution to motivated performance.

Brain levels of DA also fluctuate to reflect the preparation and energized-execution of motivated behaviors. For instance, microdialysis measurement of striatal DA levels correlate with ongoing reward rate (Onge et al., 2012), a motivational decision variable for task engagement. Additionally, several voltammetric measurements of dopamine release during reward seeking reveal transient elevations in striatal [DA] accompanying motivated approach behaviors (Phillips et al., 2003; Roitman et al., 2004a; Wassum et al., 2012; Syed et al., 2016). Together these observations indicate a role of DA (especially, mesolimbic DA) in the generation and execution of motivated reward-seeking behaviors. This notion is formalized in the *incentive salience theory* of DA, wherein Kent Berridge and colleagues (Berridge and Robinson, 1998) argue that DA plays a critical role in invigorating actions towards desired goals. This view of DA primarily underscores its contributions to immediate behaviors; i.e modulation of current performance.

The incentive-salience theory generally describes psychological and neural underpinnings of intense desires (or “wanting”) that influence the urgent production of actions to seek (or interact with) rewards and associated stimuli. In this view, the degree of vigor for approach behaviors can suddenly diverge from stored (learned) strengths of brain representations for reward. Berridge (Berridge and ODoherty, 2014) notes: “It is

possible to “want” what is not expected to be liked, nor remembered to be liked, as well as what is not actually liked when obtained”. Mechanistically, Berridge proposes that fluctuations in hormones and brain neurotransmitters underlying the current motivational state can immediately amplify (and even invert) the stored value of rewards. In one computational formalization of this account, Zhang et al (Zhang et al., 2009) propose that the motivational state value (i.e incentive value, $V(s_t)$) is dependent on the sum of discounted future rewards. Critically, however, the value of rewards can (selectively and immediately) be modulated by a kappa factor (equation 1).

$$V(s_t) = \check{r}(r_t * \kappa) + \gamma V(s_{t+1}) \quad (1)$$

This kappa factor (assumed $\kappa=1$ during training) is predicted to multiplicatively scale the immediate motivational value of rewards. Changes in relevant physiological states (such as increases in hunger or dietary sodium appetite) selectively elevate kappa above 1 for the particular reward. This will lead to amplification of desires to pursue the reward (increased decision utility) and hence emission of highly invigorated approach and consummatory behaviors. Mesolimbic DA is, thus, argued to reflect this kappa factor, and surges in striatal DA are hypothesized to lead to augmentation of incentive salience for rewards, and associated stimuli.

In support of this view, many DA manipulations interfere with the vigor of online behavioral performance. For example, pavlovian cue-evoked approach behaviors are energized by intracranial injections of amphetamines (Wyvell and Berridge, 2000; Smith et al., 2011), whereas pharmacological blockade of striatal DA receptors increases latency and decreases the probability of conditioned responses (Nicola, 2010; Saunders and Robinson, 2012; Flagel et al., 2011; Robinson and Flagel, 2009). Further, reaction times in operant tasks were also sensitive to manipulations that enhanced or blunted dopamine transmission (Leventhal et al., 2014), whereas direct optogenetic stimulation of the VTA promotes approach behaviors (Ilango et al., 2014). Finally, cells in the ventral pallidum are reported to encode the incentive salience associated with conditioned stimuli that energize behavior (Richard et al., 2016; Smith et al., 2011; Tindell et al., 2009; Ahrens et al., 2016).

In addition to DA system's response to primary rewards (such as sex, food or drugs), DA cell firing and subsequent striatal DA release are promoted by cues associated with rewards. Of note, in a series of seminal studies, Wolfram Schultz and his colleagues characterized changes in midbrain DA firing during pavlovian association of cues to rewards (Ljungberg et al., 1992; Schultz et al., 1993; Mirenowicz and Schultz, 1994; Schultz, 1997). Specifically, the trial-by-trial time-course of firing across training episodes appeared to progressively shift from rewards themselves to cues that predict rewards. These findings provided the theoretical basis for another influential theory of DA function: the *reward prediction error (RPE)* hypothesis (Schultz, 1997). This view of DA primarily underscores its contributions to future behavior by promoting learning.

Simply put, the DA-RPE hypothesis essentially argues that phasic activation (and suppression) of midbrain cells encode information about surprising (or disappointing) variation of rewards from expectation. Ideally, a fully predicted reward is neither surprising nor disappointing. But, unexpected delivery of rewards or omissions generate error signals proportional to their deviation from prediction. These error signals are hypothesized to drive learning, by providing instruction about *when* and *how strongly* to update internal expectations to reward.

This theory is grounded in temporal-difference reinforcement learning (Schultz, 2010; Bayer and Glimcher, 2005; Niv, 2009; Schultz, 1997), and posits that DA *is* the error signal used for behavioral reinforcement. Despite its notable limitations (Redgrave et al., 2010; Redgrave et al., 2008; Niv, 2013; Berridge, 2007), DA-RPE theory is the current dominant description for DA function, successfully accounting for many behavioral and neural data from humans, monkeys and rodents. Indeed, part of the reason for its success is that it integrates smoothly with normative computational frameworks for brain mechanisms of reinforcement learning.

The Reinforcement Learning framework

Reinforcement Learning (RL) methods are machine learning algorithms that describe formal computational methods for solving reward-learning problems. In this

framework, an agent continuously interacts with its environment to obtain rewards, but is never told what actions to take. The agent must discover which actions lead to rewards through trial-and-error. Key to succeeding in this framework is learning a set of policies that map actions onto experienced states, so that optimal actions are selected from candidates (Sutton and Barto, 2012; Niv, 2009; Singh et al., 2010). The agent's exposure to the environment is organized into discrete states that themselves follow transition rules. Lastly, a global reward-function dictates the schedule of reinforcement, and the algorithm executes iteratively (or continuously) to incrementally update stored values for state-action pairs.

A specific class of RL algorithms that have achieved considerable success in translation to animal behavior and neuroscience are called Actor-Critic (AC) methods (Figure 1.3). Here, an actor node stores and updates action values (for example left or right weights, wL or wR) for each trial, while a critic node maintains a separate state value (V, the expected discounted sum of rewards expected from all future states, equation 2).

$$V_t = E[\gamma^0 r_t + \gamma^1 r_{t+1} + \gamma^2 r_{t+2} \ldots] \quad (2)$$

Gamma (γ) governs the rate at which rewards in the future are discounted. Once the outcome of a trial is revealed (r_t = 1 or 0), a state value error signal (δ_t) is evaluated as the difference between expected reward (V_t) and attained reward (r_t, equation 3).

$$\delta_t = V_t - r_t \quad (3)$$

This error signal drives incremental update of both state (equation 4) and action (equation 5,6) values according to learning rates of the critic(α_{critic}) and actor (α_{actor}) nodes respectively.

$$V_{t+1} = V_t + (\alpha_{critic} * \delta_t) \quad (4)$$

$$wL_{t+1} = wL_t + (\alpha_{actor} * \delta_t) \quad (5)$$

$$wR_{t+1} = wR_t + (\alpha_{actor} * \delta_t) \quad (6)$$

Several methods allow for selecting among candidate actions on a given trial, for example epsilon greedy (pick action with largest value on (1 - ε) proportion of trials), or softmax method (log-gradient normalized transfer function).

$$pL = 1 / [1 + e^{-\beta * (wL_t - wR_t)}] \quad (7)$$

The softmax selection method (equation 7) is common in neuroscience implementation of RL methods. One reason is because an inverse temperature parameter (β) allows for the dynamic modulation of the degree to which cached action values have influence over choice. In other words, the actor's probabilistic choice can be tuned to accommodate exploitation of learned action values (bias toward favored action), or exploring options (allocate choice randomly). This exploration/exploitation tradeoff is also apparent in selection behaviors of humans and rodents (Cohen et al., 2007; Jepma and Nieuwenhuis, 2011; Frank et al., 2009; Beeler et al., 2010). I have omitted the discussion of many aspects of RL relevant to the brain, but see other reviews for more details (Niv, 2009; Ribas-Fernandes et al., 2011; Niv et al., 2005; Niv et al., 2006; Joel et al., 2002). Below, I review the broad mechanics for brain implementation in the AC model (Figure 1.4).

The overlying cortex represents the current state (defined by many combinations of environmental stimuli) and the BG nuclei implement the actor and critic modules (Frank, 2011; Cohen and Frank, 2009; Joel et al., 2002). In this view, the critic (which learns state values) is likely implemented in loops involving the orbitofrontal cortex, ventral striatum and amygdala. On the other hand the actor module learns response action-values and is hypothesized to be implemented in the dorsal striatum (Maia and Frank, 2011; Doya, 2000; Niv, 2009). Values for actions and stimuli are represented in the synaptic strength of cortical inputs to MSNs. Specifically, one variant of the AC method (Collins and Frank, 2014; Maia and Frank, 2016) predicts that positive and negative evidence for action values may respectively be stored in dMSNs and iMSN. Finally, RPE signals from midbrain DA neurons are used to update state and action

values (i.e modulate strength of cortical synapses via induction of LTP and LTD (Shen et al., 2008)) in proportion to the difference in the anticipated and observed outcomes (Schultz, 2007; Dayan, 2009; Redish, 2004).

Many empirical reports provide evidence for predictions of AC framework. For instance, similar to what has been repeatedly observed in brain slices, striatal synapses also experience robust (and bidirectional) plasticity in vivo (Reynolds et al., 2001; Xiong et al., 2015). Specifically, Xiong et al found that inputs from auditory cortex (tuned to reward-associated frequencies) undergo selective strengthening, whereas neighboring inputs coding for unrewarded tones do not (Xiong et al., 2015). In addition, endogenous dopamine release with a specific temporal coincidence to glutamate release causes MSN spine growth selectively at glutamergic synapses (Yagishita et al., 2014). Direct stimulation of MSNs affects future behaviors consistent with circuit modifications to promote learning (Gerfen and Surmeier, 2011; Maia and Frank, 2011). For example, optogenetic activation of MSNs modulate the likelihood of action re-selection on subsequent trials (respectively producing reinforcement or punishment in D1 or D2 cells) (Kravitz et al., 2012).

The phasic firing of midbrain DA neurons are correlated with positive RPEs related to the unexpected changes in magnitude, probability or timing of rewards (Mirenowicz and Schultz, 1994; Schultz et al., 1993; Bayer and Glimcher, 2005; Cohen et al., 2012; Nakahara et al., 2004; Satoh et al., 2003; Eshel et al., 2016; Schultz, 1997; Schultz, 1998; Redgrave et al., 2008). These phasic increases in DA likely promote simultaneous LTP in dMSNs and LTD in iMSNs (Shen et al., 2008), to promote associative or instrumental learning (Schultz, 2007). Consistent with this notion, optogenetic stimulation of DA cells drives associative learning (Steinberg et al., 2013), and operant conditioning (Tsai et al., 2009; Tye et al., 2013; Rossi et al., 2013). Conversely, pauses in spiking activity cause transient decreases in striatal dopamine concentration (McCutcheon et al., 2014), but whether these pauses can signal the full range of negative RPEs is debated (Bayer and Glimcher, 2005). Indeed, brief

optogenetic silencing of DA cells does weakens instrumental associations, suggesting a causal role of these pauses in learning (Chang et al., 2016).

Summary

DA transmission is key for brain computations of reward prediction, and the energization of movements to seek rewards. While it is clear that DA exerts its influence on behavior by modulating the underlying circuit for adaptive behavior, the precise role of DA continues to be debated. There are two major theories of DA function: the *incentive salience DA theory* emphasizes its role in current motivation, and the *RPE-coding hypothesis* stresses a view of DA for learning. In this thesis I aim to further our understanding of DA dynamics during adaptive decision-making, and assess *how* DA makes simultaneous contributions to learning and performance vigor.

I present three data chapters below, and each assesses a defined set of questions. Chapter two describes results from several experiments using a range of techniques directed at assessing the apparent simultaneous (and complementary) contributions of mesolimbic dopamine to learning and motivation. We report that, across multiple timescales, DA signals a single decision variable: a Value function (an ongoing estimate of discounted future rewards) that is used for motivational decision making ('Is It worth it?') and that abrupt changes in this value function serve as temporal-difference reward prediction errors used for learning ("repeat action?”). Chapter three makes use of precisely (and variably) timed optogenetic manipulation of the VTA to determine the ‘plasticity windows’ underlying state and action value learning. Finally, in chapter four, I used fiber photometry to assess if the decision-signals encoded in midbrain DA neurons reflect those of striatal DA concentrations. These experiments are motivated by studies reporting cellular mechanisms of disparity between spiking and DA release within the striatum.

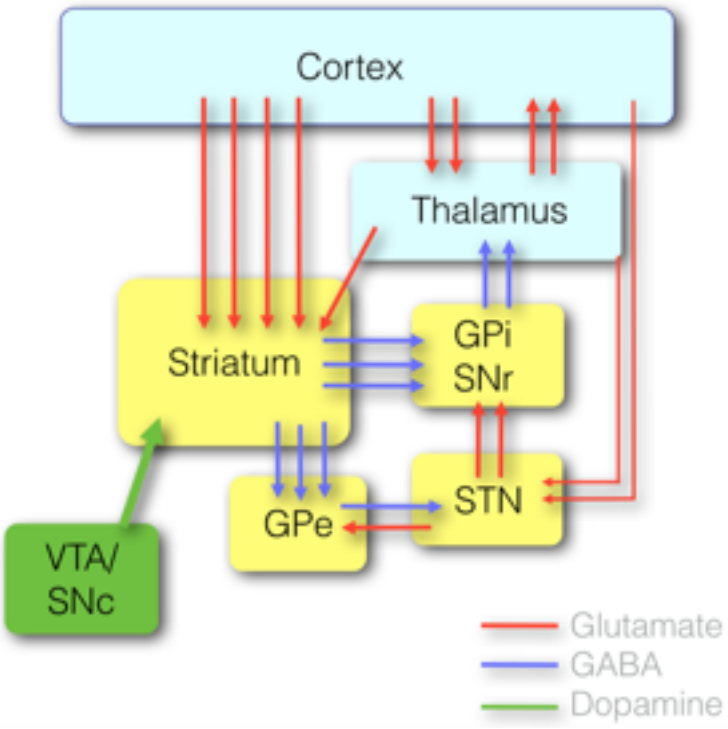

Figure 1.1: A simple schematic of BG nuclei and their connections

Here, red arrows indicate glutamergic synapses, blue are inhibitory GABAergic, and green arrow is a dopaminergic connection. Many BG nuclei make reciprocal connections, but are simplified here with unidirectional projections.

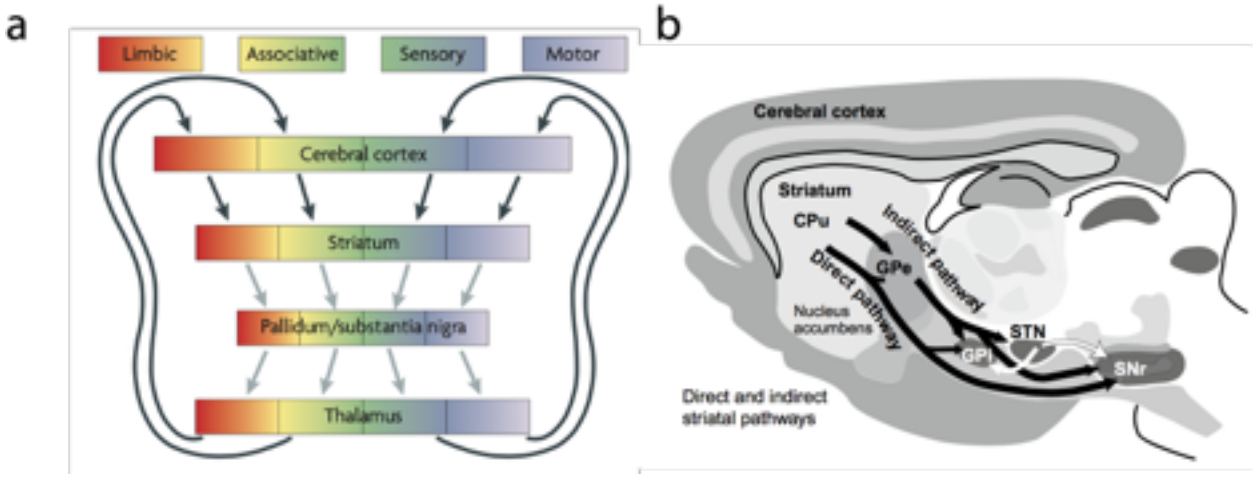

Figure 1.2: Two major schemes of BG organization:

BG Macro and micro-circuits organized into (a) repeating compiles limbic, associative, sensory and motor loops (each recruiting identical micro-circuit modules). Image reproduced from (Redgrave et al., 2010). (b) Further, BG connectivity is also segregated into direct, indirect and hyperdirect (not pictured) pathways. Image reproduced from Gerfen and Bolam, 2010

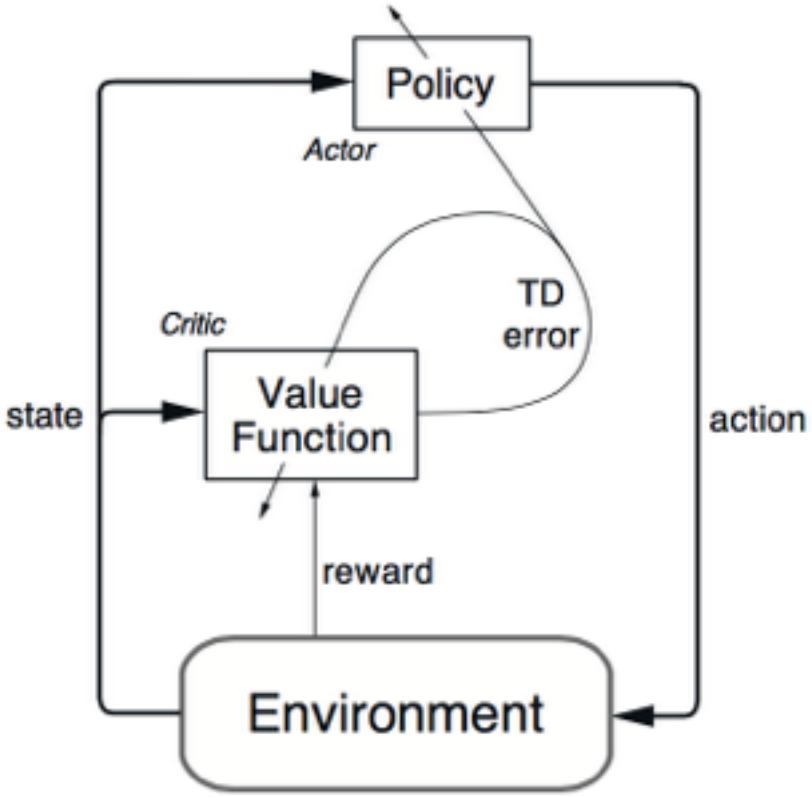

Figure 1.3: Actor-Critic components

In the actor-critic view, an agent interacts with the environment via actions, and perceives states. Predicted return from the current state is stored as value functions in the critic, and a selection policy is implemented in the actor to maximize returns. Unexpected changes in reward schedule prompts TD errors, that update predictions of reward in the critic, and action values in the actor. Image reproduced from Sutton and Barto, 2012.

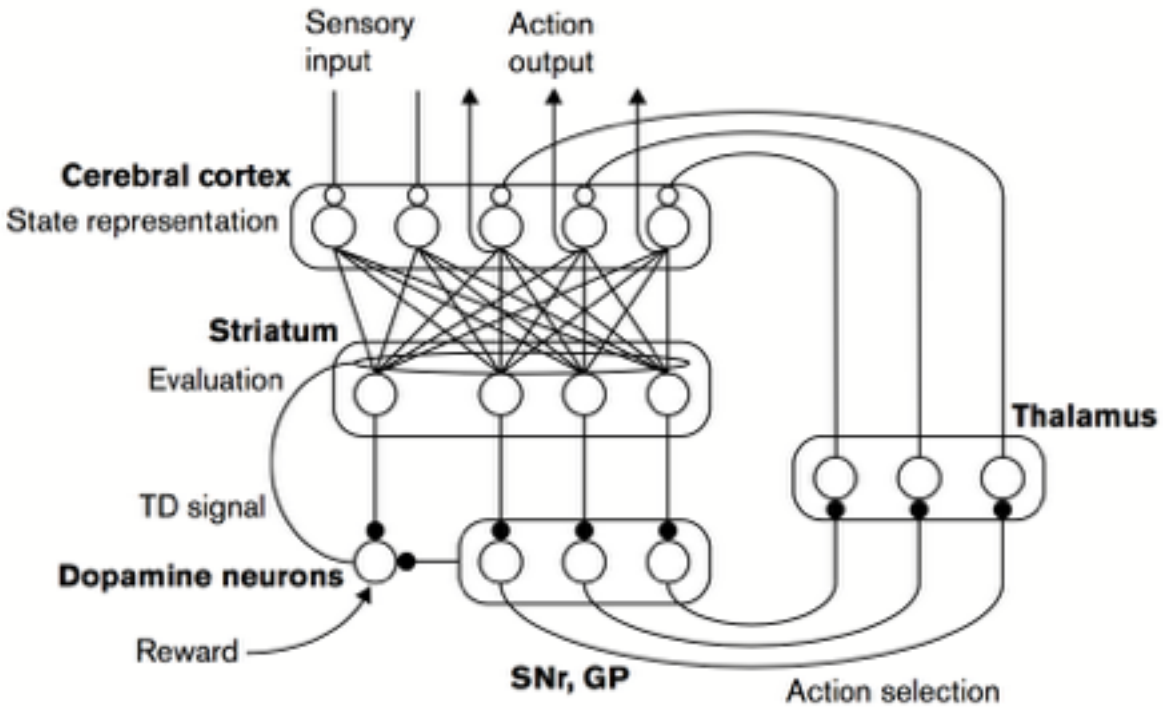

Figure 1.4: A computational framework for adaptive decision making in the BG

A schematic of information flow, and computations performed by brain areas recruited by reinforcement learning. Image reproduced from Doya, 2000.

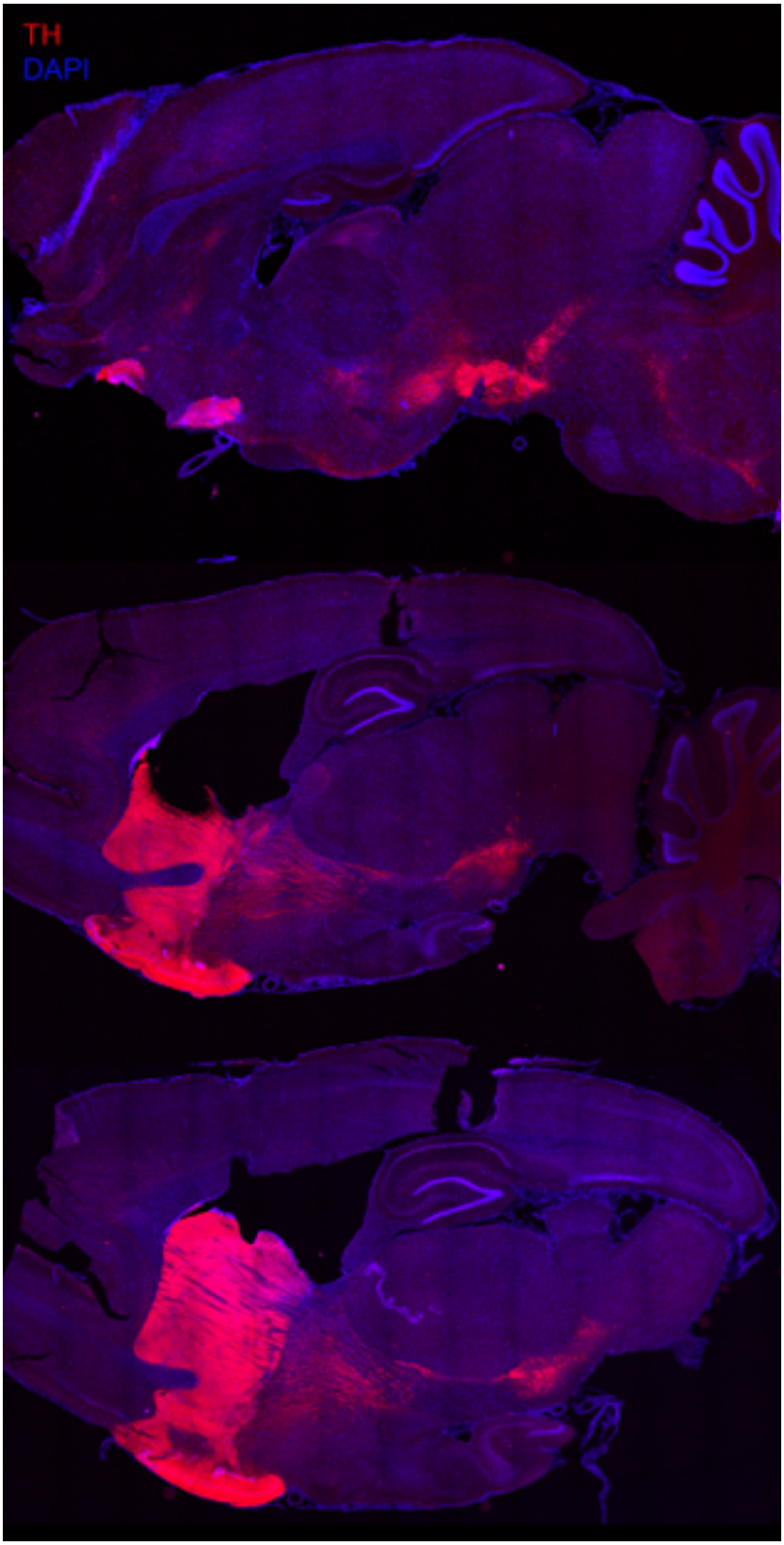

Figure 1.5: DA system of the rat brain

Antibody staining of tyrosine hydroxylase expression in sagittal sections of the brain. Top slice is most medial, and images below are progressively lateral. DA cells are clustered in the midbrain and send long range axons along the medial forebrain bundle to forebrain locations including the striatum and prefrontal cortex.

References

Ahmed, S.H., and Cador, M. (2006). Dissociation of psychomotor sensitization from compulsive cocaine consumption. Neuropsychopharmacology *31*, 563-571.

Ahrens, A.M., Meyer, P.J., Ferguson, L.M., Robinson, T.E., and Aldridge, J.W. (2016). Neural Activity in the Ventral Pallidum Encodes Variation in the Incentive Value of a Reward Cue. J Neurosci *36*, 7957-7970.

Albin, R.L., Young, A.B., and Penney, J.B. (1989). The functional anatomy of basal ganglia disorders. Trends Neurosci *12*, 366-375.

Alexander, G.E., DeLong, M.R., and Strick, P.L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci *9*, 357-381.

Amini, B., Clark, J.W., and Canavier, C.C. (1999). Calcium dynamics underlying pacemaker-like and burst firing oscillations in midbrain dopaminergic neurons: a computational study. J Neurophysiol *82*, 2249-2261.

Aragona, B.J., Day, J.J., Roitman, M.F., Cleaveland, N.A., Wightman, R.M., and Carelli, R.M. (2009). Regional specificity in the real-time development of phasic dopamine transmission patterns during acquisition of a cue-cocaine association in rats. Eur J Neurosci *30*, 1889-1899.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129-141.

Beckstead, R.M., and Kersey, K.S. (1985). Immunohistochemical demonstration of differential substance P-, met-enkephalin-, and glutamic-acid-decarboxylase-containing

cell body and axon distributions in the corpus striatum of the cat. Journal of Comparative Neurology *232*, 481-498.

Beeler, J.A., Daw, N., Frazier, C.R., and Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. Front Behav Neurosci *4*, 170.

Berendse, H.W., Graaf, Y.G.-D., and Groenewegen, H.J. (1992a). Topographical organization and relationship with ventral striatal compartments of prefrontal corticostriatal projections in the rat. Journal of Comparative Neurology *316*, 314-347.

Berendse, H.W., Groenewegen, H.J., and Lohman, A.H. (1992b). Compartmental distribution of ventral striatal neurons projecting to the mesencephalon in the rat. The Journal of neuroscience *12*, 2079-2103.

Berridge, K.C. (2007). The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) *191*, 391-431.

Berridge, K.C., and ODoherty, J.P. (2014). From experienced utility to decision utility. Neuroeconomics (Second Edition) , 335-351.

Berridge, K.C., and Robinson, T.E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Research Reviews *28*, 309 - 369.

Berridge, K.C., and Robinson, T.E. (2003). Parsing reward. Trends Neurosci *26*, 507-513.

Björklund, A., and Lindvall, O. (1984). Dopamine-containing systems in the CNS. Handbook of chemical neuroanatomy *2*, 55-122.

Cachope, R., Mateo, Y., Mathur, B.N., Irving, J., Wang, H.L., Morales, M., Lovinger, D.M., and Cheer, J.F. (2012). Selective activation of cholinergic interneurons enhances

accumbal phasic dopamine release: setting the tone for reward processing. Cell Rep *2*, 33-41.

Calabresi, P., Picconi, B., Tozzi, A., and Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends Neurosci *30*, 211-219.

Centonze, D., Grande, C., Saulle, E., Martín, A.B., Gubellini, P., Pavón, N., Pisani, A., Bernardi, G., Moratalla, R., and Calabresi, P. (2003). Distinct roles of D1 and D5 dopamine receptors in motor activity and striatal synaptic plasticity. The Journal of neuroscience *23*, 8506-8512.

Cepeda, C., Buchwald, N.A., and Levine, M.S. (1993). Neuromodulatory actions of dopamine in the neostriatum are dependent upon the excitatory amino acid receptor subtypes activated. Proceedings of the National Academy of Sciences *90*, 9576-9580.

Chang, C.Y., Esber, G.R., Marrero-Garcia, Y., Yau, H.J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci *19*, 111-116.

Cohen, J.D., McClure, S.M., and Angela, J.Y. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society of London B: Biological Sciences *362*, 933-942.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85-88.

Cohen, M.X., and Frank, M.J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. Behav Brain Res *199*, 141-156.

Collins, A.G., and Frank, M.J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychological review *121*, 337.

Cui, G., Jun, S.B., Jin, X., Pham, M.D., Vogel, S.S., Lovinger, D.M., and Costa, R.M. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. Nature *494*, 238-242.

Dahlström, A., and Fuxe, K. (1964). Evidence for the existence of monoamine-containing neurons in the central nervous system. I. Demonstration of monoamines in the cell bodies of brain stem neurons. Acta Physiologica Scandinavica. Supplementum , SUPPL-SU232.

Day, M., Wokosin, D., Plotkin, J.L., Tian, X., and Surmeier, D.J. (2008). Differential excitability and modulation of striatal medium spiny neuron dendrites. The Journal of neuroscience *28*, 11603-11614.

Dayan, P. (2009). Dopamine, reinforcement learning, and addiction. Pharmacopsychiatry *42 Suppl 1*, S56-S65.

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. Curr Opin Neurobiol *18*, 185-196.

DeLong, M.R. (1983). The neurophysiologic basis of abnormal movements in basal ganglia disorders. Neurobehavioral Toxicology & Teratology

DeLong, M.R. (1990). Primate models of movement disorders of basal ganglia origin. Trends Neurosci *13*, 281-285.

Descarries, L., Watkins, K.C., Garcia, S., Bosler, O., and Doucet, G. (1996). Dual character, asynaptic and synaptic, of the dopamine innervation in adult rat neostriatum: a quantitative autoradiographic and immunocytochemical analysis. Journal of Comparative Neurology *375*, 167-186.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. Current opinion in neurobiology *10*, 732-739.

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. Nat Neurosci *19*, 479-486.

Fallon, J.H. (1988). Topographic Organization of Ascending Dopaminergic Projectionsa. Annals of the New York Academy of Sciences *537*, 1-9.

Fiorino, D.F., Coury, A., and Phillips, A.G. (1997). Dynamic changes in nucleus accumbens dopamine efflux during the Coolidge effect in male rats. The Journal of neuroscience *17*, 4849-4855.

Flagel, S.B., Clark, J.J., Robinson, T.E., Mayo, L., Czuj, A., Willuhn, I., Akers, C.A., Clinton, S.M., Phillips, P.E., and Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. Nature *469*, 53-57.

Floresco, S.B., West, A.R., Ash, B., Moore, H., and Grace, A.A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. Nat Neurosci *6*, 968-973.

Frank, M.J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw *19*, 1120-1136.

Frank, M.J. (2011). Computational models of motivated action selection in corticostriatal circuits. Curr Opin Neurobiol *21*, 381-386.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci *12*, 1062-1068.

Freeze, B.S., Kravitz, A.V., Hammack, N., Berke, J.D., and Kreitzer, A.C. (2013). Control of basal ganglia output by direct and indirect pathway projection neurons. J Neurosci *33*, 18531-18539.

Galarraga, E., Herná-López, S., Reyes, A., Barral, J., and Bargas, J. (1997). Dopamine facilitates striatal EPSPs through an L-type Ca2+ conductance. Neuroreport *8*, 2183-2186.

Gao, T., Yatani, A., Dell'Acqua, M.L., Sako, H., Green, S.A., Dascal, N., Scott, J.D., and Hosey, M.M. (1997). cAMP-dependent regulation of cardiac L-type Ca 2+ channels requires membrane targeting of PKA and phosphorylation of channel subunits. Neuron *19*, 185-196.

Gerfen, C.R., and Bolam, J.P. (2010). The neuroanatomical organization of the basal ganglia. Handbook of basal ganglia structure and function *20*, 3-28.

Gerfen, C.R., and Surmeier, D.J. (2011). Modulation of striatal projection systems by dopamine. Annu Rev Neurosci *34*, 441-466.

Gerfen, C.R., and Young, W.S. (1988). Distribution of striatonigral and striatopallidal peptidergic neurons in both patch and matrix compartments: an in situ hybridization histochemistry and fluorescent retrograde tracing study. Brain Res *460*, 161-167.

Gerfen, C.R., Engber, T.M., Mahan, L.C., Susel, Z., Chase, T.N., Monsma Jr, F.J., and Sibley, D.R. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. Science *250*, 1429-1432.

Gerfen, C.R., Herkenham, M., and Thibault, J. (1987). The neostriatal mosaic: II. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems. The Journal of neuroscience *7*, 3915-3934.

Grace, A.A. (1991). Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. Neuroscience *41*, 1-24.

Grace, A.A., and Bunney, B.S. (1984a). The control of firing pattern in nigral dopamine neurons: burst firing. The Journal of neuroscience *4*, 2877-2890.

Grace, A.A., and Bunney, B.S. (1984b). The control of firing pattern in nigral dopamine neurons: single spike firing. The Journal of neuroscience *4*, 2866-2876.

Grace, A.A., Floresco, S.B., Goto, Y., and Lodge, D.J. (2007). Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. Trends Neurosci *30*, 220-227.

Graybiel, A.M. (1998). The basal ganglia and chunking of action repertoires. Neurobiology of learning and memory *70*, 119-136.

Haber, S.N., and Watson, S.J. (1983). The comparison between enkephalin-like and dynorphin-like immunoreactivity in both monkey and human globus pallidus and substantia nigra. Life sciences *33*, 33-36.

Haber, S.N., Fudge, J.L., and McFarland, N.R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. The Journal of neuroscience *20*, 2369-2382.

Hernández-Echeagaray, E., Starling, A.J., Cepeda, C., and Levine, M.S. (2004). Modulation of AMPA currents by D2 dopamine receptors in striatal medium-sized spiny neurons: are dendrites necessary? European Journal of Neuroscience *19*, 2455-2463.

Hernández-López, S., Tkatch, T., Perez-Garci, E., Galarraga, E., Bargas, J., Hamm, H., and Surmeier, D.J. (2000). D2 dopamine receptors in striatal medium spiny neurons reduce L-Type Ca2+ currents and excitability vía a novel PLCβ1--IP3--calcineurin-signaling cascade. The Journal of Neuroscience *20*, 8987-8995.

Higley, M.J., and Sabatini, B.L. (2010). Competitive regulation of synaptic Ca2+ influx by D2 dopamine and A2A adenosine receptors. Nat Neurosci *13*, 958-966.

Hikosaka, O., Takikawa, Y., and Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. Physiol Rev *80*, 953-978.

Ilango, A., Kesner, A.J., Broker, C.J., Wang, D.V., and Ikemoto, S. (2014). Phasic excitation of ventral tegmental dopamine neurons potentiates the initiation of conditioned approach behavior: parametric and reinforcement-schedule analyses. Front Behav Neurosci *8*, 155.

Ingham, C.A., Hood, S.H., Taggart, P., and Arbuthnott, G.W. (1998). Plasticity of synapses in the rat neostriatum after unilateral lesion of the nigrostriatal dopaminergic pathway. The Journal of neuroscience *18*, 4732-4743.

Jepma, M., and Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration--exploitation trade-off: evidence for the adaptive gain theory. Journal of cognitive neuroscience *23*, 1587-1596.

Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw *15*, 535-547.

Kebabian, J.W., and Greengard, P. (1971). Dopamine-sensitive adenyl cyclase: possible role in synaptic transmission. Science *174*, 1346-1349.

Kerr, J.N., and Wickens, J.R. (2001). Dopamine D-1/D-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro. J Neurophysiol *85*, 117-124.

Kita, H. (2007). Globus pallidus external segment. Prog Brain Res *160*, 111-133.

Kitai, S.T., and Surmeier, D.J. (1992). Cholinergic and dopaminergic modulation of potassium conductances in neostriatal neurons. Advances in neurology *60*, 40-52.

Kiyatkin, E.A., Wise, R.A., and Gratton, A. (1993). Drug-and behavior-associated changes in dopamine-related electrochemical signals during intravenous heroin self-administration in rats. Synapse *14*, 60-72.

Kravitz, A.V., Freeze, B.S., Parker, P.R., Kay, K., Thwin, M.T., Deisseroth, K., and Kreitzer, A.C. (2010). Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. Nature *466*, 622-626.

Kravitz, A.V., Tye, L.D., and Kreitzer, A.C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat Neurosci *15*, 816-818.

Kreitzer, A.C. (2009). Physiology and pharmacology of striatal neurons. Annu Rev Neurosci *32*, 127-147.

Kreitzer, A.C., and Malenka, R.C. (2007). Endocannabinoid-mediated rescue of striatal LTD and motor deficits in Parkinson's disease models. Nature *445*, 643-647.

Leventhal, D.K., Stoetzner, C.R., Abraham, R., Pettibone, J., DeMarco, K., and Berke, J.D. (2014). Dissociable effects of dopamine on learning and performance within sensorimotor striatum. Basal Ganglia *4*, 43-54.

Lindvall, O., Björklund, A., Moore, R.Y., and Stenevi, U. (1974). Mesencephalic dopamine neurons projecting to neocortex. Brain Res *81*, 325-331.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol *67*, 145-163.

Maia, T.V., and Frank, M.J. (2011). From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci *14*, 154-162.

Maia, T.V., and Frank, M.J. (2016). An Integrative Perspective on the Role of Dopamine in Schizophrenia. Biol Psychiatry

Matsuda, W., Furuta, T., Nakamura, K.C., Hioki, H., Fujiyama, F., Arai, R., and Kaneko, T. (2009). Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum. J Neurosci *29*, 444-453.

McCutcheon, J.E., Cone, J.J., Sinon, C.G., Fortin, S.M., Kantak, P.A., Witten, I.B., Deisseroth, K., Stuber, G.D., and Roitman, M.F. (2014). Optical suppression of drug-evoked phasic dopamine release. Front Neural Circuits *8*

Middleton, F. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. Brain Research Reviews *31*, 236-250.

Mink, J.W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. Progress in neurobiology *50*, 381-425.

Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. J Neurophysiol *72*, 1024-1027.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. Neuron *41*, 269-280.

Nelson, A.B., and Kreitzer, A.C. (2014). Reassessing models of basal ganglia function and dysfunction. Annu Rev Neurosci *37*, 117-135.

Nicola, S.M. (2010). The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. J Neurosci *30*, 16585-16600.

Niv, Y. (2009). Reinforcement learning in the brain. Journal of Mathematical Psychology *53*, 139-154.

Niv, Y. (2013). Neuroscience: Dopamine ramps up. Nature *500*, 533-535.

Niv, Y., Duff, M.O., and Dayan, P. (2005). Dopamine, uncertainty and TD learning. Behavioral and Brain Functions *1*, 1-9.

Niv, Y., Joel, D., and Dayan, P. (2006). A normative perspective on motivation. Trends in cognitive sciences *10*, 375-381.

Onge, J.R.S., Ahn, S., Phillips, A.G., and Floresco, S.B. (2012). Dynamic fluctuations in dopamine efflux in the prefrontal cortex and nucleus accumbens during risk-based decision making. The Journal of Neuroscience *32*, 16880-16891.

O'Reilly, R.C., and Frank, M.J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput *18*, 283-328.

Pawlak, V., and Kerr, J.N. (2008). Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. The Journal of Neuroscience *28*, 2435-2446.

Penney, J.B., and Young, A.B. (1983). Speculations on the functional anatomy of basal ganglia disorders. Annu Rev Neurosci *6*, 73-94.

Penny, G.R., Afsharpour, S., and Kitai, S.T. (1986). The glutamic acid decarboxylase-, leucine-, enkephalin-, and substance P-immunoreactive neurons in the neostriatum of the rat and cat: evidence for partial population overlap. Neuroscience *17*, l-1045.

Pereira, D.B., Schmitz, Y., Mészáros, J., Merchant, P., Hu, G., Li, S., Henke, A., Lizardi-Ortiz, J.E., Karpowicz, R.J., Morgenstern, T.J., Sonders, M.S., Kanter, E., Rodriguez, P.C., Mosharov, E.V., Sames, D., and Sulzer, D. (2016). Fluorescent false neurotransmitter reveals functionally silent dopamine vesicle clusters in the striatum. Nat Neurosci

Phillips, P.E., Stuber, G.D., Heien, M.L., Wightman, R.M., and Carelli, R.M. (2003). Subsecond dopamine release promotes cocaine seeking. Nature *422*, 614-618.

Rebec, G.V., Christensen, J.R., Guerra, C., and Bardo, M.T. (1997). Regional and temporal differences in real-time dopamine efflux in the nucleus accumbens during free-choice novelty. Brain Res *776*, 61-67.

Redgrave, P., Coizet, V., and Reynolds, J. (2010). Phasic dopamine signaling and basal ganglia function. Handbook of basal ganglia structure and function , 549-559.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? Brain Res Rev *58*, 322-339.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M.R., and Obeso, J.A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. Nat Rev Neurosci *11*, 760-772.

Redish, A.D. (2004). Addiction as a computational process gone awry. Science *306*, 1944-1947.

Reynolds, J.N., Hyland, B.I., and Wickens, J.R. (2001). A cellular mechanism of reward-related learning. Nature *413*, 67-70.

Ribas-Fernandes, J., Solway, A., Diuk, C., McGuire, J., Barto, A., Niv, Y., and Botvinick, M. (2011). A Neural Signature of Hierarchical Reinforcement Learning. Neuron *71*, 370-379.

Rice, M.E., and Cragg, S.J. (2004). Nicotine amplifies reward-related dopamine signals in striatum. Nat Neurosci *7*, 583-584.

Richard, J.M., Ambroggi, F., Janak, P.H., and Fields, H.L. (2016). Ventral Pallidum Neurons Encode Incentive Value and Promote Cue-Elicited Instrumental Actions. Neuron *90*, 1165-1173.

Richfield, E.K., Penney, J.B., and Young, A.B. (1989). Anatomical and affinity state comparisons between dopamine D 1 and D 2 receptors in the rat central nervous system. Neuroscience *30*, 767-777.

Robinson, D.L., Phillips, P.E., Budygin, E.A., Trafton, B.J., Garris, P.A., and Wightman, R.M. (2001). Sub-second changes in accumbal dopamine during sexual behavior in male rats. Neuroreport *12*, 2549-2552.

Robinson, T.E., and Flagel, S.B. (2009). Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. Biol Psychiatry *65*, 869-873.

Roitman, M.F., Stuber, G.D., Phillips, P.E., Wightman, R.M., and Carelli, R.M. (2004a). Dopamine operates as a subsecond modulator of food seeking. J Neurosci *24*, 1265-1271.

Roitman, M.F., Stuber, G.D., Phillips, P.E.M., Wightman, R.M., and Carelli, R.M. (2004b). Dopamine Operates as a Subsecond Modulator of Food Seeking. J Neurosci *24*, 1265-1271.

Rolls, E.T., Burton, M.J., and Mora, F. (1980). Neurophysiological analysis of brain-stimulation reward in the monkey. Brain Res *194*, 339-357.

Roseberry, T.K., Lee, A.M., Lalive, A.L., Wilbrecht, L., Bonci, A., and Kreitzer, A.C. (2016). Cell-Type-Specific Control of Brainstem Locomotor Circuits by Basal Ganglia. Cell *164*, 526-537.

Rossi, M.A., Sukharnikova, T., Hayrapetyan, V.Y., Yang, L., and Yin, H.H. (2013). Operant self-stimulation of dopamine neurons in the substantia nigra. PLoS One *8*, e65799.

Salamone, J., and Correa, M. (2012). The Mysterious Motivational Functions of Mesolimbic Dopamine. Neuron *76*, 470-485.

Salamone, J.D., Correa, M., Farrar, A.M., Nunes, E.J., and Pardo, M. (2009). Dopamine, behavioral economics, and effort. Front Behav Neurosci *3*, 13.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. Science *310*, 1337-1340.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci *23*, 9913-9923.

Saunders, B.T., and Robinson, T.E. (2012). The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses. Eur J Neurosci *36*, 2521-2532.

Schultz, W. (1997). A Neural Substrate of Prediction and Reward. Science *275*, 1593-1599.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J Neurophysiol *80*, 1-27.

Schultz, W. (2007). Multiple dopamine functions at different time courses. Annu Rev Neurosci *30*, 259-288.

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. Behavioral and brain functions *6*, 1.

Schultz, W. (2015). Neuronal Reward and Decision Signals: From Theories to Data. Physiol Rev *95*, 853-951.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci *13*, 900-913.

Shen, W., Flajolet, M., Greengard, P., and Surmeier, D.J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. Science *321*, 848-851.

Shou, M., Ferrario, C.R., Schultz, K.N., Robinson, T.E., and Kennedy, R.T. (2006). Monitoring dopamine in vivo by microdialysis sampling and on-line CE-laser-induced fluorescence. Analytical chemistry *78*, 6717-6725.

Singh, S., Lewis, R.L., Barto, A.G., and Sorg, J. (2010). Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. IEEE Transactions on Autonomous Mental Development *2*, 70-82.

Smith, K.S., Berridge, K.C., and Aldridge, J.W. (2011). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. Proc Natl Acad Sci U S A *108*, E255-E264.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci *16*, 966-973.

Sulzer, D. (2011). How addictive drugs disrupt presynaptic dopamine neurotransmission. Neuron *69*, 628-649.

Sulzer, D., Zhang, H., Benoit-Marand, M., and Gonon, F. (2010). Regulation of extracellular dopamine: release and reuptake. The Handbook of Basal Ganglia Structure and Function , 297-319.

Surmeier, D.J., Bargas, J., Hemmings, H.C., Nairn, A.C., and Greengard, P. (1995). Modulation of calcium currents by a D1 dopaminergic protein kinase/phosphatase cascade in rat neostriatal neurons. Neuron *14*, 385-397.

Surmeier, D.J., Song, W.J., and Yan, Z. (1996). Coordinated expression of dopamine receptors in neostriatal medium spiny neurons. J Neurosci *16*, 6579-6591.

Sutton, R.S., and Barto, A.G. (2012). Reinforcement learning: An introduction (Cambridge Univ Press).

Sweatt, J.D. (2004). Mitogen-activated protein kinases in synaptic plasticity and memory. Current opinion in neurobiology *14*, 311-317.

Syed, E.C., Grima, L.L., Magill, P.J., Bogacz, R., Brown, P., and Walton, M.E. (2016). Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat Neurosci *19*, 34-36.

Tai, L.H., Lee, A.M., Benavidez, N., Bonci, A., and Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci *15*, 1281-1289.

Tecuapetla, F., Jin, X., Lima, S.Q., and Costa, R.M. (2016). Complementary Contributions of Striatal Projection Pathways to Action Initiation and Execution. Cell *166*, 703-715.

Threlfell, S., Lalic, T., Platt, N.J., Jennings, K.A., Deisseroth, K., and Cragg, S.J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron *75*, 58-64.

Tindell, A.J., Smith, K.S., Berridge, K.C., and Aldridge, J.W. (2009). Dynamic computation of incentive salience: "wanting" what was never "liked". J Neurosci *29*, 12220-12228.

Tsai, H.C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., and Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science *324*, 1080-1084.

Tye, K.M., Mirzabekov, J.J., Warden, M.R., Ferenczi, E.A., Tsai, H.C., Finkelstein, J., Kim, S.Y., Adhikari, A., Thompson, K.R., Andalman, A.S., Gunaydin, L.A., Witten, I.B., and Deisseroth, K. (2013). Dopamine neurons modulate neural encoding and expression of depression-related behaviour. Nature *493*, 537-541.

Voorn, P., Vanderschuren, L.J., Groenewegen, H.J., Robbins, T.W., and Pennartz, C.M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. Trends Neurosci *27*, 468-474.

Wang, Z., Kai, L., Day, M., Ronesi, J., Yin, H.H., Ding, J., Tkatch, T., Lovinger, D.M., and Surmeier, D.J. (2006). Dopaminergic control of corticostriatal long-term synaptic depression in medium spiny neurons is mediated by cholinergic interneurons. Neuron *50*, 443-452.

Wassum, K.M., Ostlund, S.B., and Maidment, N.T. (2012). Phasic mesolimbic dopamine signaling precedes and predicts performance of a self-initiated action sequence task. Biol Psychiatry *71*, 846-854.

Wightman, R.M., Heien, M.L., Wassum, K.M., Sombers, L.A., Aragona, B.J., Khan, A.S., Ariansen, J.L., Cheer, J.F., Phillips, P.E., and Carelli, R.M. (2007). Dopamine release is heterogeneous within microenvironments of the rat nucleus accumbens. Eur J Neurosci *26*, 2046-2054.

Wise, R.A. (1982). Neuroleptics and operant behavior: the anhedonia hypothesis. Behavioral and brain sciences *5*, 39-53.

Wise, R.A. (1996). Addictive drugs and brain stimulation reward. Annu Rev Neurosci *19*, 319-340.

Wise, R.A. (2004). Dopamine, learning and motivation. Nat Rev Neurosci *5*, 483-494.

Wise, R.A., and Bozarth, M.A. (1987). A psychomotor stimulant theory of addiction. Psychological review *94*, 469.

Witten, I.B., Steinberg, E.E., Lee, S.Y., Davidson, T.J., Zalocusky, K.A., Brodsky, M., Yizhar, O., Cho, S.L., Gong, S., Ramakrishnan, C., Stuber, G.D., Tye, K.M., Janak, P.H., and Deisseroth, K. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron *72*, 721-733.

Wyvell, C.L., and Berridge, K.C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward "wanting" without enhanced "liking" or response reinforcement. J Neurosci *20*, 8122-8130.

Xiong, Q., Znamenskiy, P., and Zador, A.M. (2015). Selective corticostriatal plasticity during acquisition of an auditory discrimination task. Nature

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science *345*, 1616-1620.

Yttri, E.A., and Dudman, J.T. (2016). Opponent and bidirectional control of movement velocity in the basal ganglia. Nature *533*, 402-406.

Zhang, J., Berridge, K.C., Tindell, A.J., Smith, K.S., and Aldridge, J.W. (2009). A neural computational model of incentive salience. PLoS Comput Biol *5*, e1000437.

Zhou, Q.-Y., and Palmiter, R.D. (1995). Dopamine-deficient mice are severely hypoactive, adipsic, and aphagic. Cell *83*, 1197-1209.

CHAPTER 2: MESOLIMBIC DOPAMINE SIGNALS THE VALUE OF WORK

Abstract

Dopamine cell firing can encode errors in reward prediction, providing a learning signal to guide future behavior. Yet dopamine is also a key modulator of motivation, invigorating current behavior. Existing theories propose that fast ("phasic") dopamine fluctuations support learning, while much slower ("tonic") dopamine changes are involved in motivation. We examined dopamine release in the nucleus accumbens across multiple time scales, using complementary microdialysis and voltammetric methods during adaptive decision-making. We first show that minute-by-minute dopamine levels covary with reward rate and motivational vigor. We then show that second-by-second dopamine release encodes an estimate of temporally-discounted future reward (a value function). We demonstrate that changing dopamine immediately alters willingness to work, and reinforces preceding action choices by encoding temporal-difference reward prediction errors. Our results indicate that dopamine conveys a single, rapidly-evolving decision variable, the available reward for investment of effort, that is employed for both learning and motivational functions.

* This chapter was published as:

Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. Nat Neurosci *19*, 117-126.

Introduction

Altered dopamine signaling is critically involved in many human disorders from Parkinson's Disease to drug addiction. Yet the normal functions of dopamine have long been the subject of debate. There is extensive evidence that dopamine affects learning, especially the reinforcement of actions that produce desirable results (Reynolds et al., 2001). Specifically, electrophysiological studies suggest that bursts and pauses of dopamine cell firing encode the reward prediction errors (RPEs) of reinforcement learning theory (RL) (Schultz, 1997). In this framework RPE signals are used to update estimated values of states and actions, and these updated values affect subsequent decisions when similar situations are re-encountered. Further support for a link between phasic dopamine and RPE comes from measurements of dopamine release using fast-scan cyclic voltammetry (FSCV) (Day et al., 2007; Hart et al., 2014) and optogenetic manipulations (Gan et al., 2010; Steinberg et al., 2013).

There is also extensive evidence that dopamine modulates arousal and motivation (Berridge, 2007; Beierholm et al., 2013). Drugs that produce prolonged increases in dopamine release (e.g. amphetamines) can profoundly enhance psychomotor activation, while drugs or toxins that interfere with dopamine transmission have the opposite effect. Over slow timescales (tens of minutes) microdialysis studies have demonstrated that dopamine release ([DA]) is strongly correlated with behavioral activity, especially in the nucleus accumbens (Freed and Yamamoto, 1985) (i.e. mesolimbic [DA]). It is widely thought that slow (tonic) [DA] changes are involved in motivation (Niv et al., 2007; Cagniard et al., 2006; Salamone and Correa, 2012). However, faster [DA] changes also appear to have a motivational function (Satoh et al., 2003). Subsecond increases in mesolimbic [DA] accompany motivated approach behaviors (Phillips et al., 2003; Roitman et al., 2004), and dopamine ramps lasting several seconds have been reported as rats approach anticipated rewards (Howe et al., 2013), without any obvious connection to RPE. Overall, the role of dopamine in motivation is still considered "mysterious" (Salamone and Correa, 2012).

The purpose of this study was to better understand just how dopamine contributes to motivation, and to learning, simultaneously. We demonstrate that mesolimbic [DA] conveys a motivational signal in the form of *state values*, which are moment-by-moment estimates of available future reward. These values are used for making decisions about whether to *work*, i.e. to invest time and effort in activities that are not immediately rewarded, in order to obtain future rewards. When there is an unexpected change in value, the corresponding change in [DA] not only influences motivation to work, but also serves as an RPE learning signal, reinforcing specific choices. Rather than separate functions of “phasic” and “tonic” [DA], our data support a unified view in which the same dynamically-fluctuating [DA] signal influences both current and future motivated behavior.

Methods

Animals and Behavioral Task

All animal procedures were approved by the University of Michigan Committee on Use and Care of Animals. Male rats (300-500g, either wild-type Long-Evans or *TH-Cre*$^{+}$ with a Long-Evans background (Witten et al., 2011) were maintained on a reverse 12:12 light:dark cycle and tested during the dark phase. Rats were mildly food deprived, receiving 15g of standard laboratory rat chow daily in addition to food rewards earned during task performance. Training and testing was performed in computer-controlled Med Associates operant chambers (25cm x 30cm at widest point) each with a 5-hole nose-poke wall, as previously described (Gage et al., 2010; Leventhal et al., 2012; Schmidt et al., 2013). Training to perform the trial-and-error task typically took ~2 months, and included several pretraining stages (2 days-2 weeks each, advancing when ~85% of trials were performed without procedural errors). First, any one of the 5 nosepoke holes was illuminated (at random), and poking this hole caused delivery of a 45 mg fruit punch flavored sucrose pellet into the Food Port (FR1 schedule). Activation of the food hopper to deliver the pellet caused a audible click (the reward cue). In the next stage, the hole illuminated at trial start was always one of the three more-central holes (randomly-selected), and rats learned to poke and maintain hold for a variable interval (750-1250ms) until Go cue onset (250ms duration white noise, together with dimming of the start port). Next, Go cue onset was also paired with illumination of both adjacent side ports. A leftward or rightward poke to one of these ports was required to receive a reward (each at 50% probability), and initiated the inter-trial interval (5-10s randomly selected from a uniform distribution). If the rat poked an unlit center port (wrong start) or pulled out before the end of the hold period (false start), the house light turned on for the duration of an inter-trial-interval. During this stage (only), to discourage development of a side bias, a maximum of three consecutive pokes to the same side were rewarded. Finally, in the complete trial-and-error task left and right choices had independent reward probabilities, each maintained for blocks of 40-60 trials (randomly selected block length and sequence for each session). All combinations of 10%, 50%

and 90% reward probability were used except 10:10 and 90:90. There was no event that indicated to the rat that a trial would be unrewarded other than the omission of the Reward cue and the absence of the pellet.

For a subset of ChR2 optogenetic sessions, overhead video was captured at 15 frames/s. The frames immediately preceding the Light-On events were extracted, and the positions of the nose tip and neck were marked (by scorers blind to whether that trial included laser stimulation). These positions were used to determine rat distance and orientation to the center port (the one that will be illuminated on that trial). Each trial was classified as “engaged” or “unengaged”, using cutoff values of distance (10.6cm) and orientation (84°) that minimized the overlap between aggregate pink and green distributions. To assess how path length was affected by optogenetic stimulation, rat head positions were scored for each video frame between Light-On and Center-Nose-In. Engaged trials were further classified by whether the rat was immediately adjacent to one of the three possible center ports, and if that port was the one that became illuminated at Light-On or not (i.e. lucky, unlucky guesses).

Smoothing of latency (and other) time series for graphical display (Figure 2.1B,C) was performed using the MATLAB *filtfilt* function with a 7-trial window. To quantify the impact of prior trial rewards on current trial latency, we used a multiple regression model:

$$\log_{10}(latency) = \beta_1 r_{t-1} + \beta_2 r_{t-2} + \ldots \beta_{10} r_{t-10}$$

where r = 1 if the corresponding trial was rewarded. All latency analyses excluded trials of zero latency (i.e. those for which the rat’s nose was already inside the randomly-chosen center port at Light-On). For analysis of prior trial outcomes on left/right choice behavior we used another multiple regression model, just as previously described (Lau and Glimcher, 2005).

Latency survivor curves were calculated simply as the proportion of trials for which the Center-In event had not yet occurred, at each 250ms interval after Light-On (i.e. a inverted cumulative latency distribution), smoothed with a 3-point moving average ($x_t' = 0.25x_{t-1} + 0.5x_t + 0.25x_{t+1}$). These survivor curves were then used to calculate hazard rates, as the fraction of the remaining latencies that occurred in each 250ms bin (i.e. the number of Center-In events that happened, divided by the number that could have happened).

We defined reward rate as the exponentially-weighted moving average of individual rewards (i.e., a leaky integrator (Simen et al., 2006; Daw et al., 2002; Sugrue et al., 2004)). For each session the integrator time constant was chosen to maximize the (negative) correlation between reward rate and behavioral latency. If instead we defined reward rate as simply the number of rewards during each minute (i.e. ignoring the contributions of trials in previous minutes to current reward rate), the relationship between microdialysis-measured [DA] in that minute and reward rate was lower, though still significant (R^2=0.084, $p=5.5x10^{-10}$).

An important parameter in reinforcement learning is the degree to which agents choose the option that is currently estimated to be the best (exploitation) versus trying alternatives to assess whether they are actually better (exploration), and dopamine has been proposed to mediate this trade-off (Humphries et al., 2012; Beeler et al., 2012). To assess this we examined left/right choices in the second half of each block, by which time choices have typically stabilized (Fig. 1D; this behavioral pattern was also seen for the microdialysis sessions). We defined an Exploitation Index as the proportion of trials for which rats choose the better option in these second block halves (so values close to 1 would be fully exploitative, and values close to 0.5 would be random/exploratory). As an alternative metric of exploration/exploitation, we examined the number of times that the rat switched between left and right choices in each minute; this metric also showed no significant relationship to any neurochemical assayed in our microdialysis experiments.

Microdialysis

After 3-6 months of behavioral training rats were implanted with guide cannulae bilaterally above the nucleus accumbens core (NAcc; +1.3-1.9mm AP, 1.5mm ML from bregma) and allowed to recover for at least one week before retraining. On test days (3-5 weeks after cannula implantation) a single custom made microdialysis probe (300μm diameter) with polyacrylonitrile membrane (Hospal, Bologna, Italy; 20kD molecular weight cutoff) was inserted into NAcc, extending 1mm below the guide cannula. Artificial CSF (composition in mM: $CaCl_2$ 1.2; KCl 2.7, NaCl 148, $MgCl_2$ 0.85, 0.25 ascorbate) was perfused continuously at 2μl/min. Rats were placed in the operant chamber with the house light on for an initial 90min period of probe equilibration, after which samples were collected once every minute. Following five baseline samples the house light was extinguished to indicate task availability.

For chemical analyses, we employed a modified version of our benzoyl chloride derivatization and HPLC-MS analysis method (Song et al., 2011). Immediately after each 2μl sample collection, we added 1.5μl of buffer (sodium carbonate monohydrate 100 mM), 1.5μl of 2% benzoyl chloride in acetonitrile, and 1.5μl of a 13C-labeled internal standard mixture (total mixture volume 6.5μl). The mixture was vortexed for 2s between each reagent addition. Since ACh is a quaternary amine and thus not derivatized by benzoyl chloride, it was directly detected in its native form (transition 146->87). Deuterated ACh (d4-ACh) was also added to the internal standard mixture for improved ACh quantification (Song et al., 2011). Five μl of the sample mixture was automatically injected by a Thermo Accela HPLC system (Thermo Fisher Scientific, Waltham, MA) onto a reverse-phase Kinetex biphenyl HPLC column (2.1mm x 100 mm; 1.7 particle size; Phenomenex, Torrance CA). The HPLC system was interfaced to a HESI II ESI probe and Thermo TSQ Quantum Ultra (Thermo Scientific) triple quadrupole mass spectrometer operating in positive mode. Sample run times for all analytes were 3 min. To quantify neurochemicals in dialysate samples, we constructed 6-point external calibration curves encompassing known physiological concentrations. Thermo Xcalibur 2.1 software (Thermo Fisher Scientific) automatically detected

chromatographic peaks and quantified concentrations. To reduce noise each resulting minute-by-minute time series was smoothed with a 3-point moving average (as above), then converted to Z-scores to facilitate comparison between subjects.

Regression analysis of microdialysis data was performed stepwise. We first constructed models with only one behavioral variable as predictor and one outcome (analyte). If two behavioral variables showed a significant relationship to a given analyte, we constructed a model with both behavioral variables and an interaction term, and examined the capacity of each variable to explain analyte variance without substantial multicollinearity.

To determine cross-correlogram statistical thresholds we first shuffled the time series for all sessions 200,000 times, and calculated the average Pearson correlation coefficients (i.e. the zero-lag cross-correlation) for each shuffled pair of time series. Thresholds were based on the tails of the resulting distribution: i.e. for uncorrected two-tailed alpha=0.05 we would find the levels for which 2.5% of the shuffled values lay outside these thresholds. As we wished to correct for multiple comparisons we divided alpha by the number of tests (276; number of cross-correlograms = 23 timeseries * 22 timeseries divided by two, since the crosscorrelograms are just mirror-reversed when the order is changed, plus 23 autocorrelograms).

Voltammetry

FSCV electrode construction, data acquisition and analysis were performed as described (Aragona et al., 2009). Rats were implanted with a guide cannula above the right NAcc (+1.3-2.0 mm AP, 1.5 mm ML from bregma), a Ag/AgCl reference electrode (in the contralateral hemisphere) and a bipolar stimulation electrode aimed at the VTA (-5.2 mm AP, 0.8 mm ML, 7.5 mm DV). Carbon fiber electrodes were lowered acutely into the NAcc. Dopaminergic current was quantified offline using principal component regression (PCR) (Heien et al., 2005) using training data for dopamine and pH from electrical stimulations. Recording time points that exceeded the PCR residual analysis threshold (Qα) were omitted from further processing or analysis. Current to [DA]

conversion was based on *in vitro* calibrations of electrodes constructed in the same manner with the same exposed fiber length. On many days data was not recorded due to electrode breakage or obvious movement-related electrical noise. FSCV recordings were made from 41 sessions (14 rats total). We excluded those sessions for which the rat failed to complete at least three blocks of trials, and those in which electrical artifacts caused >10% of trials to violate the assumptions of PCR residual analysis. The remaining 10 sessions came from 6 different rats. To avoid aggregate results being overly skewed by a single animal, we only included one session from each of the six rats (the session with the largest reward-evoked [DA] increase). Upon completion of FSCV testing, animals were deeply anesthetized and electrolytic lesions were created (40 μA for 15 seconds at the same depth as recording site) using stainless steel electrodes with 500 μm of exposed tip (AM Systems, USA). Lesion locations were later reconstructed in Nissl stained sections.

For between-session comparisons we normalized [DA] to the average [DA] difference between the pre-trial baseline and Food-Port-In aligned peak levels. To visualize the reward-history-dependence of [DA] change between consecutive trials (Figure 2.5H), we first extracted time series of normalized [DA] from consecutive pairs of rewarded trials (Side-In event to subsequent Side-In event separated by less than 30s). For each session we divided these traces into “low-reward-rate” and “high-reward-rate” groups, using the (# of rewarded trials in the last 10) that best approximated a median-split (i.e. so low- and high- reward-rate groups had similar trial numbers). We then averaged all low-reward-rate traces, and separately all high-reward-rate traces.

Reinforcement Learning model

To estimate the time-varying state value and RPE within each trial we used a Semi-Markov Decision Process (Daw et al., 2006) with temporal difference learning, implemented in MATLAB. The model consisted of a set of states, with rat behavioral events determining the times of transitions between states (Figure 2.11). Each state

was associated with a stored ('cached') value of entering that state, V(s). At each state transition a reward prediction error δ was calculated using:

$$\delta_t = r_t + V_t(s_t) - \gamma^{-n} V_t(s_{t-n})$$

where n is the number of timesteps since the last state transition (a timestep of 50ms was used throughout), r is defined as one at reward receipt and zero otherwise, and ɣ specifies the rate at which future rewards are discounted at each timestep (ɣ < 1). The V terms in the equation compare the cached value of the new state to the value predicted, given the prior state value and the elapsed time since the last transition (as illustrated in Figure 2.4C). Each state also had e(s), an eligibility trace that decayed with the same time parameter ɣ (following the terminology of ref. 21, this is a TD(1) model with replacing traces). RPEs updated the values of the states encountered up to that point, using:

$$V'(s) = V(s) + \alpha.\delta.e_t(s)$$

where α is the learning rate. V and ɣ were defined only at state transitions, and V was constrained to be non-negative. The model was "episodic" as all eligibilities were reset to zero at trial outcome (reward receipt, or omission). V is therefore a estimate of the time-discounted value of the next reward, rather than total aggregate future reward; with exponential discounting and best-fit parameters subsequent time-discounted rewards are negligible (but this would not necessarily be the case if hyperbolic discounting was used).

We also examined the effect of calculating prediction errors slightly differently:

$$\delta_t = r_t + \gamma^{n} V_t(s_t) - V_t(s_{t-n})$$

This version compares a discounted version of the new state value to the previous state value. As expected, the results were the same. Specifically, overall [DA] correlation to V remained ~0.4, overall δ correlation was ~0.2, and each individual session [DA] was significantly better correlated to V than to δ, across the full parameter space.

We present results using ɣ in the 0.9 to 1 range, because 0.9 is already a very fast exponential discount rate when using 50ms timesteps. However we also tested smaller ɣ (0.05-0.9) and confirmed that the [DA]: δ correlation only diminished in this lower range (not shown).

To compare within-trial [DA] changes to model variables, we identified all epochs of time (3s before to 3s after Center-In) with at least 6 state transitions (this encompasses both rewarded and unrewarded trials). Since the model can change state value instantaneously, but our FSCV signal cannot (Kile et al., 2012), we included an offset lag (so we actually compared V and δ to [DA] a few measurements later). The size of the lag affected the magnitude of the observed correlations (Figure 2.4f) but not the basic result. Results were also unchanged if (instead of a lag) we convolved model variables with a kernel consisting of an exponential rise and fall (Figure 2.12), demonstrating that our results are not a simple artifact of time delays associated with the FSCV method or sluggish reuptake. Finally, we also tried using the SMDP model with hyperbolic (instead of exponential) discounting (Mazur, 1984; Ainslie, 2005; Kobayashi and Schultz, 2008; Kacelnik, 1997), and again found a consistently stronger correlation between [DA] and V than between [DA] and δ (not shown).

Optogenetics

We used three groups of rats to assess the behavioral effects of VTA DA cell manipulations (first *TH-Cre*$^{+}$ with AAV-EF1α-DIO-ChR2-EYFP virus, then littermate *TH-Cre*$^{-}$ with the same virus, then *TH-Cre*$^{+}$ with AAV-EF1α-DIO-eNpHR3.0-EYFP). All virus was produced at the University of North Carolina vector core. In each case rats received bilateral viral injections (0.5 or 1μl per hemisphere at 50 nL/min) into the VTA (same

coordinates as above). After 3 weeks, we placed bilateral optic fibers (200 μm diameter) under ketamine/xylazine anesthesia with FSCV guidance, at an angle of 6° from the sagittal plane, stopping at a location that yielded the most laser-evoked [DA] release in NAc. Once cemented in place, we used FSCV to test multiple sets of stimulation parameters from a 445nm blue laser diode (Casio) with Arroyo Instruments driver under LabView control. The parameters chosen for behavioral experiments (0.5s train of 10ms pulses at 30Hz, 20mW power at tip) typically produced [DA] increases in *TH-Cre*$^{+}$ / ChR2 rats comparable to those seen with unexpected reward delivery. All rats were allowed to recover from surgery and retrained to pre-surgery performance. Combined behavioral / optogenetic experiments began 5 weeks after virus injection. On alternate days, sessions either included bilateral laser stimulation (on a randomly selected 30% of trials, regardless of block or outcome), or not. In this manner, each rat received 3 sessions of Light-On stimulations and 3 sessions of Side-In stimulation, interleaved with control (no laser) sessions, over a two-week period. Halorhodopsin rats were tested with 1s of constant 20mW illumination from a 589nm (yellow/orange) laser (OEM Systems), starting either at Light-On or Side-In as above. One *TH-Cre*$^{+}$ /ChR2 rat was excluded from analyses due to misplaced virus (no viral expression directly below the optic fiber tips).

For statistical analysis of optogenetic effects on behavior we used repeated measure ANOVA models, in SPSS. For each rat we first averaged data across the 3 sessions with the same optogenetic conditions. Then, to assess reinforcing effects we examined the two factors of LASER (off vs on) and REWARD (rewarded vs omission), with the dependent measure the probability that the same action was repeated on the next trial. For assessing effects on median latency we examined the two factors of LASER (off vs on) and ENGAGED (yes vs no). For assessing group-dependent effects on hazard rate we examined the factors of LASER (off vs on) and GROUP (*TH-Cre*$^{+}$ / ChR2; *TH-Cre*$^{-}$ /ChR2; *TH-Cre*$^{+}$ /eNpHR3.0), with the dependent measure the average hazard rate during the epoch 1-2.5s after Light-On. This epoch was chosen since it is 1-2s after the laser stimulation period (0-0.5s) and approach behaviors have a

consistent duration of ~1-2s (Figure 2.15). Post-hoc tests were Bonferroni-corrected for multiple comparisons.

Results

Motivation to work adapts to recent reward history

We made use of an adaptive decision-making task (Figure 2.1a and Methods) that is closely related to the reinforcement learning framework (a “two-armed bandit”). On each trial a randomly-chosen nose poke port lit up (Light-On) indicating that the rat might profitably approach and place its nose in that port (Center-In). The rat had to wait in this position for a variable delay (0.75-1.25s), until an auditory white noise burst (Go cue) prompted the rat to make a brief leftward or rightward movement to an adjacent side port. Unlike previous behavioral tasks using the same apparatus, the Go cue did not specify which way to move; instead the rat had to learn through trial-and-error which option was currently more likely to be rewarded. Left and right choices had separate reward probabilities (each either 10%, 50%, or 90%), and these probabilities changed periodically without any explicit signal. On rewarded trials only, entry into the side port (Side-In) immediately triggered an audible click (the reward cue) as a food hopper delivered a sugar pellet to a separate food port at the opposite side of the chamber.

Trained rats readily adapted their behavior in at least two respects (Figure 2.1b,c). Firstly, actions followed by rewards were more likely to be subsequently selected (i.e. they were reinforced), producing left/right choice probabilities that scaled with actual reward probabilities (Samejima et al., 2005) (Figure 2.1d).

Secondly, rats were more motivated to perform the task while it was producing a higher rate of reward (Guitart-Masip et al., 2011; Wang et al., 2013). This was apparent from “latency” (the time taken from Light-On until the Center-In nose poke), which scaled inversely with reward rate (Figure 2.1e-g). When reward rate was higher rats were more likely to be already waiting near the center ports at Light-On (“engaged” trials; Figure 2.7), producing very short latencies. Higher reward rates also produced shorter latencies even when rats were not already engaged at Light-On (Figure 2.7), due to an elevated moment-by-moment probability (hazard rate) of choosing to begin work (Figure 2.1h,i).

These latency observations are consistent with optimal foraging theories (Stephens and Krebs, 1986), which argue that reward rate is a key decision variable ("currency"). As animals perform actions and experience rewards they construct estimates of reward rate, and can use these estimates to help decide whether engaging in an activity is worthwhile. In a stable environment, the best estimate of reward rate is simply the total magnitude of past rewards received over a long time period, divided by the duration of that period. It has been proposed that such a "long-term-average reward rate" is encoded by slow (tonic) changes in [DA] (Niv et al., 2007). However, under shifting conditions such as our trial-and-error task, the reward rate at a given time is better estimated by more local measures. Reinforcement learning algorithms use past reward experiences to update estimates of future reward from each state: a set of these estimates is called a value function (Sutton and Barto, 1998).

Minute-by-minute dopamine correlates with reward rate

To test whether changes in [DA] accompany reward rate during adaptive decision-making, we first employed microdialysis in the nucleus accumbens combined with liquid chromatography - mass spectrometry. This method allows us to simultaneously assay a wide range of neurochemicals, including all of the well-known low-molecular weight striatal neurotransmitters, neuromodulators and their metabolites (Figure 2.2a), each with 1-minute time resolution. We performed regression analyses to assess relationships between these neurochemicals and a range of behavioral factors: 1) reward rate; 2) the number of trials attempted (as an index of a more general form of activation/arousal); 3) the degree of exploitation versus exploration (an important decision parameter that has been suggested to involve [DA]; see Methods), and 4) the cumulative reward obtained (as an index of progressively increasing factors such as satiety).

We found a clear overall relationship between [DA] and ongoing reward rate ($R^2 = 0.15$, $p < 10^{-16}$). Among the 19 tested analytes, [DA] had by far the strongest relationship to reward rate (Figure 2.2b), and this relationship was significant in 6 of 7

individual sessions, from 6 different rats (Figure 2.8). Modest significant relationships were also found for the dopamine metabolites DOPAC and 3-MT. We found a weaker relationship between [DA] and the number of trials attempted, but this was entirely accounted for by reward rate - i.e. if the regression model already included reward rate, adding number of attempts did not improve model fit. We did not find support for alternative proposals that tonic [DA] is related to exploration or exploitation, since higher [DA] was not associated with an altered probability of choosing the better left/right option (Figure 2.2b, Figure 2.8). [DA] also showed no relationship to the cumulative total rewards earned (though there was a strong relationship between cumulative reward and the dopamine metabolite HVA, among other neurochemicals; Figure 2.2b). Additional information about the relationships between neurochemicals and behavioral variables, and one another, is given in Figure 2.9.

We conclude that higher reward rate is associated specifically with higher average [DA], rather than other striatal neuromodulators, and with increased motivation to work. This finding supports the proposal that [DA] helps to mediate the effects of reward rate on motivation (Niv et al., 2007). However, rather than signaling an especially “long-term” rate of reward, [DA] tracked minute-by-minute fluctuations in reward rate. We therefore needed to assess whether this result truly reflects an aspect of [DA] signaling that is inherently slow (tonic), or could instead be explained by rapidly-changing [DA] levels, that signal a rapidly-changing decision variable.

Dopamine signals time-discounted available future reward

To help distinguish these possibilities we used FSCV to assess task-related [DA] changes on fast time scales (from tenths of seconds to tens of seconds; Figure 2.3). Within each trial, [DA] rapidly increased as rats poked their nose in the start hole (Figure 2.3c,d, Center-In panel) and for all rats this increase was more closely related to this approach behavior than to the onset of the light cue (for data from each of the single sessions from all 6 rats, see Figure 2.10). A second abrupt increase in [DA] occurred following presentation of the Go cue (Figure 2.3c,d, middle panel). If received, the

reward cue prompted a third abrupt increase (Figure 2.3c,d, Side-In panel). [DA] rose still further as rat approached the food port (Figure 2.3c,d, right), then declined once the reward was obtained. The same overall pattern of task-related [DA] change was observed in all rats, albeit with some variation (Figure 2.10). [DA] increases did not simply accompany movements, since on the infrequent trials in which the rat approached the food port without hearing the reward cue we observed no corresponding increase in [DA] (Figure 2.3c,d, right).

The overall ramping up of [DA] as rats drew progressively closer to reward suggested some form of reward expectation (Howe et al., 2013). Specifically, we hypothesized that [DA] continuously signals a *value function*: the temporally-discounted reward predicted from the current moment. To make this more clear, consider a hypothetical agent moving through a sequence of distinct, unrewarded states leading up to an expected reward (Figure 2.4a, top; perhaps a rat running at constant speed along a familiar maze arm). Since the reward is more discounted when more distant, the value function will progressively rise until the reward is obtained.

This value function describes the time-varying level of motivation. If a reward is distant (so strongly discounted), animals are less likely to choose to work for it. Once engaged however, animals are increasingly motivated, and so less likely to quit, as they detect progress towards the reward (the value function produces a “goal-gradient”, in the terminology of Hull (Hull, 1932)). If the reward is smaller or less reliable, the value function will be lower, indicating less incentive to begin work. Moving closer to our real situation, suppose that reward is equally likely to be obtained, or not, on any given trial, but a cue indicates this outcome halfway through the trial (Figure 2.4a, bottom). The increasing value function should initially reflect the overall 0.5 reward probability, but if the reward cue occurs estimated value should promptly jump to that of the (discounted) full reward.

Such unpredicted sudden transitions to states with a different value produce “temporal-difference” RPEs (Figure 2.4b). In particular, if the value function is low (e.g. the trajectory indicating 0.25 expectation of reward), the reward cue produces a large

RPE, as value jumps up to the discounted value of the now-certain reward. If instead reward expectation was higher (e.g. 0.75 trajectory), the RPE produced by the reward cue is smaller. Since temporal difference RPEs *are* rapid shifts in value, under some conditions they can be challenging to dissociate from value itself. However, RPE and value signals are not identical. In particular, as reward gets closer, the state value progressively increases but RPE remains zero unless events occur with unpredicted value or timing.

Our task includes additional features, such as variable timing between events and many trials. We therefore considered what the "true" value function should look like - on average - based on actual times to future rewards (Figure 2.4c). At the beginning of a trial, reward is at least several seconds away, and may not occur at all until a later trial. During correct trial performance each subsequent, variably-timed event indicates to the rat that rewards are getting closer and more likely, and thus causes a jump in state value. For example, hearing the Go cue indicates both that reward is closer, and that the rat will not lose out by moving too soon (an impulsive procedural error). Hearing the reward cue indicates that reward is now certain, and only a couple of seconds away.

To assess how the intertwined decision variables - state value and RPE - are encoded by phasic [DA], we compared our FSCV measurements to the dynamically varying state value and RPE of a reinforcement learning model (see Methods). This simplified model consisted of a set of discrete states (Figure 2.11), whose values were updated using temporal-difference RPEs. When given as input the actual sequence of behavioral events experienced by the rat, the model's value function consisted of a series of increases within each trial (Fig. 4d,e), resembling the observed time course of [DA] (Figure 2.3c).

Consistent with the idea that state value represents motivation to work, model state value early in each trial significantly correlated with behavioral latencies for all rats (across a wide range of model parameter settings; Figure 2.11). We identified model parameters (learning rate = 0.4, discount factor = 0.95) that maximized this behavioral correlation across all rats combined, and examined the corresponding within-trial

correlation between [DA] and model variables. For all of the 6 FSCV rats we found a clear and highly significant positive correlation between phasic [DA] and state value V (Fig. 4f). [DA] and RPE were also positively correlated, as expected since V and RPE partially covary. However, in every case [DA] had a significantly stronger relationship to V than to RPE (Figure 2.4f, Figure 2.11). We emphasize that this result was *not* dependent on specific model parameters; in fact, even if parameters were chosen to maximize the [DA] : RPE correlation, the [DA] : V correlation was stronger (Figure 2.11).

Correlations were maximal when V was compared to the [DA] signal measured ~0.4-0.5s later (Figure 2.4g). This small delay is consistent with the known brief lag associated with the FSCV method using acute electrodes (Venton et al., 2002), and prior observations that peak [DA] response occurs ~0.5s after cue onset with acute FSCV recordings (Day et al., 2007). As an alternative method of incorporating temporal distortion that might be produced by FSCV and/or the finite speeds of DA release and update, we convolved model variables with a kernel consisting of an exponential rise and fall, and explored the effect of varying kernel time constants. Once again [DA] always correlated much better with V than with RPE, across a wide range of parameter values (Figure 2.12). We conclude that state value provides a more accurate description of the time course of [DA] fluctuations than RPE alone, even though RPEs can be simultaneously signaled as changes in state value.

Abrupt dopamine changes encode reward prediction errors.

FSCV electrode signals tend to drift over a timescale of minutes, so standard practice is to assess [DA] fluctuations relative to a pre-trial “baseline” of unknown concentration (as in Figure 2.3). Presented this way, reward cues appeared to evoke a higher absolute [DA] level when rewards were less common (Figure 2.5a,b), consistent with a conventional RPE-based account of phasic [DA]. However, our model implies a different interpretation of this data (Fig. 4b, 5c). Rather than a jump from a fixed to a variable [DA] level (that encodes RPE), we predicted that the reward cue actually causes a [DA] jump from a variable [DA] level (reflecting variable estimates of upcoming

reward) to a fixed [DA] level (that encodes the time-discounted value of the now certain reward).

To test these competing accounts, we compared [DA] levels between consecutive pairs of rewarded trials with Side-In events < 30s apart (i.e. well within the accepted stable range of FSCV measurements (Heien et al., 2005); for included pairs of trials the average time between Side-In events was 11.5s). If the [DA] level evoked by the reward cue reflects RPE, then this level should tend to decline as rats experience consecutive rewards (Figure 2.5d,e). However, if [DA] represents state value then "baseline" [DA] should asymptotically increase with repeated rewards while reward cue-evoked [DA] remains more stable (Figure 2.5f,g). The latter proved correct (Figure 2.5h,i). These results provide clear further evidence that [DA] reflects reward expectation (the value function), not just RPE.

Considering the microdialysis and FSCV results together, a parsimonious interpretation is that, across multiple measurement time scales, [DA] simply signals estimated availability of reward. The higher minute-by-minute [DA] levels observed with greater reward rate reflect both the higher values of states distal to rewards (including "baseline" periods between active trial performance) and the greater proportion of time spent in high-value states proximal to rewards.

By conveying an estimate of available reward, mesolimbic [DA] could be used as a motivational signal, helping to decide whether it is worthwhile to engage in effortful activity. At the same time, abrupt *relative* changes in [DA] could be detected and used as an RPE signal for learning. But is the brain actually using [DA] to signal motivation, or learning, or both, within this task?

Dopamine both enhances motivation and reinforces choices

To address this question we turned to precisely-timed, bidirectional, optogenetic manipulations of dopamine. Following an approach validated in previous studies

(Steinberg et al., 2013), we expressed channelrhodopsin-2 (ChR2) selectively in dopamine neurons by combining *TH-Cre*$^{+}$ rats with DIO-ChR2 virus injections and bilateral optic fibers in the ventral tegmental area (Figure 2.13). We chose optical stimulation parameters (10ms pulses of blue light at 30Hz, 0.5s total duration; Figure 2.6a,b) that produced phasic [DA] increases of similar duration and magnitude to those naturally observed with unexpected reward delivery. We provided this stimulation at one of two distinct moments during task performance. We hypothesized that enhancing [DA] coincident with Light-On would increase the estimated motivational value of task performance; this would make the rat more likely to initiate an approach, leading to shorter latencies on the same trial. We further hypothesized that enhancing [DA] at the time of the major RPE (Side-In) would affect learning, as reflected in altered behavior on subsequent trials. In each session stimulation was given at only one of these two times, and on only 30% of trials (randomly selected) to allow within-session comparisons between stimulated and unstimulated trials.

Providing phasic [DA] at Side-In reinforced choice behavior: it increased the chance that the same left or right action was repeated on the next trial, whether or not the food reward was actually received (Figure 2.6c, left; n=6 rats; two-way ANOVA yielded significant main effects for LASER, $F(1,5)=224.0$, $p=2.4\times10^{-5}$ and for REWARD, $F(1,5)=41.0$, $p=0.0014$, without a significant LASER * REWARD interaction; see also Figure 2.14c). No reinforcing effect was seen if the same optogenetic stimulation was given in littermate controls (Figure 2.6c middle; n=6 *TH-Cre*$^{-}$ rats; LASER main effect $F(1,5)=2.51$, $p=0.174$). For a further group of *TH-Cre*$^{+}$ animals (n=5) we instead used the inhibitory opsin Halorhodopsin (eNpHR3.0). Inhibition of dopamine cells at Side-In reduced the probability that the same left/right choice was repeated on the next trial (LASER main effect $F(1,4)=18.7$, $p=0.012$, without a significant LASER * REWARD interaction). A direct comparison between these three rat groups also demonstrated a group-specific effect of Side-In laser stimulation on choice reinforcement (two-way ANOVA, LASER * GROUP interaction $F(2,14)=69.4$, $p=5.4\times10^{-8}$). These observations support the hypothesis that abrupt [DA] fluctuations serve as an RPE learning signal, consistent with prior optogenetic manipulations[7]. However, extra [DA] at Side-In did not

affect subsequent trial latency (Figure 2.14a,b), indicating that our artificial [DA] manipulations reproduced some, but not all, types of behavioral change normally evoked by rewarded trials.

Optogenetic effects on reinforcement were temporally-specific: providing extra [DA] at Light-On (instead of Side-In) on trial *n* did not affect the probability that rats made the same choice on trial *n+1* (LASER main effect $F(1,5) = 0.031$, $p = 0.867$; see also Figure 2.14c) nor did it affect the probability that choice on trial *n* was the same as trial *n-1* (LASER main effect $F(1,5) = 0.233$, $p=0.649$).

By contrast, extra [DA] at Light-On dramatically affected latency for that very same trial (Fig. 6d, S8). The effect on latencies depended on what the rat was doing at the time of Light-On (two-way ANOVA yielded a significant LASER * ENGAGED interaction, $F(1,3) = 28.1$, $p=0.013$). If the rat was already engaged in task performance, the very short latencies became slightly longer on average (median control latency = 0.45s, median stimulated latency=0.61s; simple main effect of LASER, $F(1,3) = 10.4$, $p = 0.048$). This effect apparently resulted from additional laser-evoked orienting movements on a subset of trials (see Figure 2.15 for more detailed analysis). By contrast, for non-engaged trials extra [DA] significantly reduced latencies (Fig. 6d; median control latency=2.64s, median stimulated latency=2.16s; simple main effect of LASER, $F(1,3) = 32.5$, $p=0.011$). These optogenetic results are consistent with the idea that mesolimbic [DA] is less important for the initiation of simple, cue-evoked responses when a task is already underway, but is critical for motivating “flexible approach” behaviors.

The shorter latencies produced by extra [DA] was not the result of rats approaching the Center-In port at faster speeds, since the average approach trajectory was unaffected (Figure 2.15). Instead, extra [DA] transiently increased the probability that rats initiated the approach behavior. As the approach itself lasted ~1-2s (Figure 2.15), the result was an increased rate of Center-In events ~1-2s after the laser pulse train (Fig. 6e; see Figure 2.70 for hazard rate time courses in individual rats). This effect of Light-On laser stimulation on hazard rates was dependent on rat group (two-way

ANOVA, LASER * GROUP interaction $F(2,14) = 26.28$, $p = 0.000018$). Post-hoc pairwise comparison of simple laser effects showed a significant increase in hazard rate for *TH-Cre+* / ChR2 rats ($F(1,14) = 62.06$, $p = 1.63 \times 10^{-6}$) and a significant reduction in hazard rate for *TH-Cre+* / eNpHR3.0 rats ($F(1,14) = 6.31$, $p = 0.025$), with no significant change in *TH-Cre-* / ChR2 rats ($F(1,14) = 2.81$, $p = 0.116$). Overall we conclude that, beyond just correlating with estimates of reward availability, mesolimbic [DA] helps translate those estimates into decisions to work for reward.

Discussion

A dopamine value signal used for both motivation and learning

Our results help confirm a range of disparate prior ideas, while placing them within a newly integrated theoretical context. First, phasic [DA] has been previously related to motivated approach (Phillips et al., 2003; Roitman et al., 2004) reward expectation (Howe et al., 2013) and effort-based decision-making (Gan et al., 2010), but our demonstration that [DA] specifically conveys the temporally-discounted value of future rewards grounds this motivational aspect of dopamine fluctuations within the quantitative frameworks of machine learning and optimal foraging theory. This idea is also consistent with findings using other techniques - for example, fMRI signals in ventral striatum (often argued to reflect dopamine signaling) encode reward expectation in the form of temporally-discounted subjective value (Kable and Glimcher, 2007).

Second, using the complementary method of microdialysis to assess slower changes we partly confirmed proposals that reward rate is reflected specifically in increased [DA], which in turn enhances motivational vigor (Niv et al., 2007). However, our critical, novel argument is that this motivational message of reward availability can dynamically change from moment to moment, rather than being an inherently slow (tonic) signal. Using optogenetics we confirmed that phasic changes in [DA] levels immediately affect willingness to engage in work, supporting the idea that sub-second [DA] fluctuations promptly influence motivational decision-making (Satoh et al., 2003; Adamantidis et al., 2011). This dynamic [DA] motivation signal can help account for detailed patterns of time allocation (Niyogi et al., 2014). For example, animals take time to reengage in task performance after getting a reward (the “post-reinforcement pause”), and this pause is longer when the next reward is smaller or more distant. This behavioral phenomenon has been a long-standing puzzle (Schlinger et al., 2008) but fits well with our argument that the time-discounted value of future rewards, conveyed by [DA], influences the moment-by-moment probability (hazard rate) of engaging in work.

Third, we confirmed the vital role of fast [DA] fluctuations, including transient dips, in signaling RPEs to affect learning (Hart et al., 2014; Kim et al., 2012; Steinberg et al., 2013). However, a striking result from our analyses is that RPEs are conveyed by fast *relative* changes in the [DA] value signal, rather than deviations from a steady (tonic) baseline. This interpretation explains for the first time how [DA] can simultaneously provide both learning and motivational signals, an important gap in prior theorizing. Our results also highlight the importance of not assuming a consistent "baseline" [DA] level across trials in voltammetry studies.

One interesting implication is that among the many postsynaptic mechanisms that are affected by dopamine, some are concerned more with absolute levels and others with fast relative changes. This possibility needs to be investigated further, together with the natural working hypothesis that [DA] effects on neuronal excitability are closely involved in motivational functions while [DA] effects on spike-timing-dependent-plasticity are responsible for reinforcement-driven learning (Reynolds et al., 2001). It is also intriguing that a pulse of increased [DA] sufficient to immediately affect latency, or to alter left/right choice on subsequent trials, does not appear sufficient to alter latency on subsequent trials. This suggests that state values and left/right action values (Samejima et al., 2005) may be updated via distinct mechanisms, or at different times within the trial.

Though dopamine is often labeled a "reward" transmitter, [DA] levels dropped during reward consumption, consistent with findings that dopamine is relatively less important for consuming - and apparently enjoying - rewards (Berridge, 2007; Cannon and Palmiter, 2003). Mesolimbic [DA] has also been shown not to be required for performance of simple actions that are immediately followed by reward, such as pressing a lever once to obtain food (Ishiwari et al., 2004). Rather, loss of mesolimbic [DA] reduces motivation to work, in the sense of investing time and effort in activities that are not inherently rewarding or interesting, but may eventually lead to rewards (Salamone and Correa, 2012). Conversely, increasing [DA] with drugs such as amphetamines increases motivation to engage in prolonged work, in both normal

subjects and those with attention-deficit hyperactivity disorder (Rapoport et al., 1980; Wardle et al., 2011).

Dopamine and decision dynamics

Our interpretation of mesolimbic [DA] as signaling the value of work is based upon rat decisions to perform our task rather than alternative "default" behaviors, such as grooming or local exploration. In this view mesolimbic [DA] helps determine *whether* to work, but not *which* activity is most worthwhile (i.e. it is "activational" more than "directional" (Salamone and Correa, 2012)). It may be best considered to signal the overall motivational excitement associated with reward expectation, or equivalently, the perceived opportunity cost of sloth (Niv et al., 2007; Niyogi et al., 2014).

Based on prior results (Gan et al., 2010) we expect that [DA] signals reward availability without factoring in the costs of effortful work, but we did not parametrically vary such costs here. Other notable limitations of this study are that we only examined [DA] in the nucleus accumbens, and we did not selectively manipulate [DA] within various striatal subregions (and other dopamine targets). Our functional account of [DA] effects on behavioral performance is undoubtedly incomplete, and it will be important to explore alternative descriptions, especially more generalizable accounts that apply throughout the striatum. In particular, our observation that mesolimbic [DA] affects the hazard rate of decisions to work seems compatible with a broader influence of striatal [DA] over decision-making - for example, by setting "thresholds" for decision process completion (Gan et al., 2010; Nagano-Saito et al., 2012; Leventhal et al., 2014). Within sensorimotor striatum dopamine influences the vigor (and learning) of more elemental actions (Leventhal et al., 2014; Haith et al., 2012), and it has been shown that even saccade speed in humans is best predicted by a discounting model that optimizes the rate of reward. In this way the activational / invigorating role of [DA] on both simple movements and motivation may reflect the same fundamental, computational-level mechanism applied to decision-making processes throughout striatum, affecting behaviors across a range of timescales.

Activational signals are useful, but not sufficient for adaptive decision-making in general. Choosing between alternative, simultaneously available courses of action requires net value representations for the specific competing options (Gan et al., 2010; Morris et al., 2006). Although different subpopulations of dopamine neurons may carry somewhat distinct signals (Matsumoto and Hikosaka, 2009), the aggregate [DA] message received by target regions is unlikely to have sufficient spatial resolution to represent multiple competing values simultaneously (Dreyer et al., 2010) or sufficient temporal resolution to present them for rapid serial consideration (McClure et al., 2003). By contrast, distinct ensembles of GABAergic neurons within the basal ganglia can dynamically encode the value of specific options, including through ramps-to-reward (Tachibana and Hikosaka, 2012; van Der Meer and Redish, 2011) that may reflect escalating bids for behavioral control. Such neurons are modulated by dopamine, and in turn provide key feedback inputs to dopamine cells that may contribute to the escalating [DA] patterns observed here.

Relationship between dopamine cell firing and release

Firing rates of presumed dopamine cells have been previously reported to escalate within trials under some conditions (Fiorillo et al., 2003), but this has not been typically reported with reward anticipation. Several factors may contribute to this apparent discrepancy with our [DA] measures. The first is the nature of the behavioral task. Many important prior studies of dopamine (Schultz, 1997; Day et al., 2007) (though not all (Morris et al., 2006)) used Pavlovian situations, in which outcomes are not determined by the animal's actions. When effortful work is not required to obtain rewards, the learned value of work may be low and corresponding decision variables may be less apparent.

Secondly, a moving rat receives constantly-changing sensory input, and may thus more easily define and discriminate a set of discrete states leading up to reward, compared to situations in which elapsed time is the sole cue of progress. When such a sequence of states can be more readily recognized, it may be easier to assign a

corresponding set of escalating values as reward gets nearer in time. Determining subjects' internal state representations, and their development during training, is an important challenge for future work. It has been argued that ramps in [DA] might actually reflect RPE if space is non-linearly represented (Gershman, 2014), or if learned values rapidly decay in time (Morita and Kato, 2014). However, these suggestions do not address the critical relationship between [DA] and motivation that we aim to account for here.

Finally, release from dopamine terminals is strongly influenced by local microcircuit mechanisms within striatum (Threlfell et al., 2012) producing a dissociation between dopamine cell firing and [DA] in target regions. This dissociation is not complete - the ability of unexpected sensory events to drive a rapid, synchronized burst of dopamine cell firing is still likely to be of particular importance for abrupt RPE signaling at state transitions. More detailed models of dopamine release, incorporating dopamine cell firing, local terminal control, and uptake dynamics will certainly be needed to understand to how [DA] comes to convey a value signal.

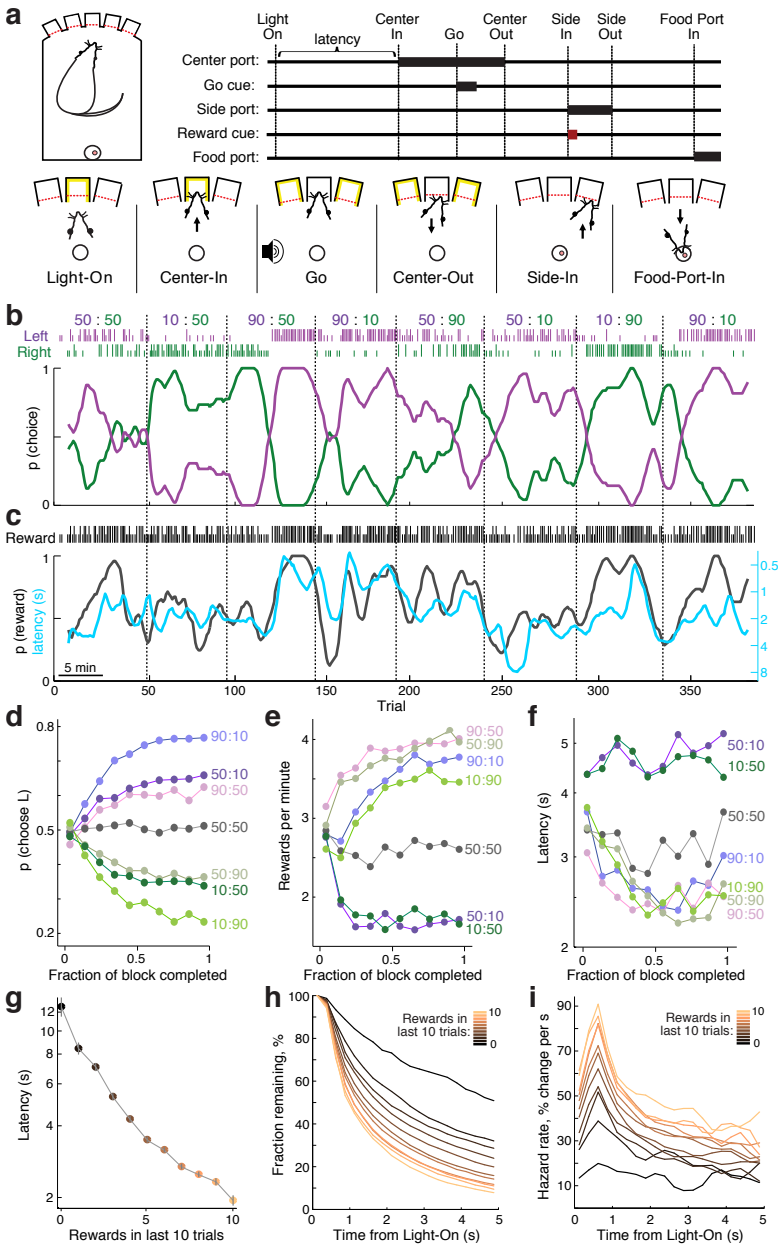

Figure 2.1: Adaptive choice and motivation in the trial-and-error task.

(a) Sequence of behavioral events (in rewarded trials). (b) Choice behavior in a representative session. Numbers at top denote nominal block-by-block reward probabilities for left (purple) and right (green) choices. Tick marks indicate actual choices and outcomes on each trial (tall ticks indicate rewarded trials, short ticks

unrewarded). The same choice data is shown below in smoothed form (thick lines; 7-trial smoothing). (c) Relationship between reward rate and latency for the same session. Here tick marks are used to indicate only whether trials were rewarded or not, regardless of choice. Solid black line shows reward rate, and cyan line shows latency (on inverted log scale), both smoothed in the same way as B. (d) Choices progressively adapt towards the block reward probabilities (data set for panels d-i: n = 14 rats, 125 sessions, 2738 +/- 284 trials per rat). (e) Reward rate breakdown by block reward probabilities. (f) Latencies by block reward probabilities. Latencies become rapidly shorter when reward rate is higher. (g) Latencies by proportion of recent trials rewarded. Error bars represent s.e.m. (h) Latency distributions presented as survivor curves (i.e. the average fraction of trials for which the Center-In event has not yet happened, by time elapsed from Light-On) broken down by proportion of recent trials rewarded. (i) Same latency distributions as panel h, but presented as hazard rates (i.e. the instantaneous probability that the Center-In event will happen, if it has not happened yet). The initial bump in the first second after Light-On reflects engaged trials (see Figure 2.7), after that hazard rates are relatively stable and continue to scale with reward history.

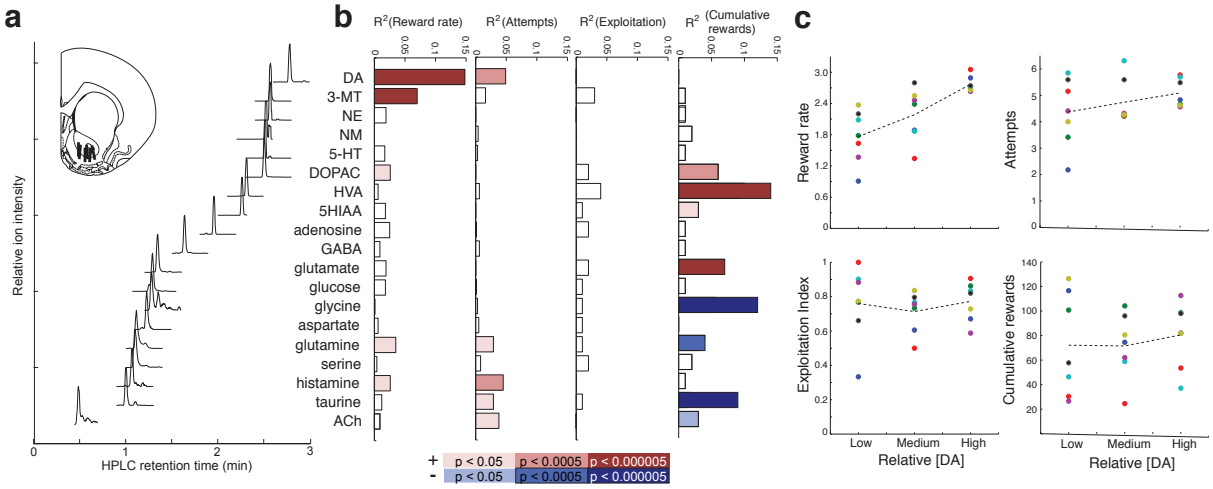

Figure 2.2: Minute-by-minute dopamine levels track reward rate.

(a) Total ion chromatogram of a single representative microdialysis sample, illustrating the set of detected analytes in this experiment. X-axis indicates HPLC retention times, y-axis indicates intensity of ion detection for each analyte (normalized to peak values). (Inset) locations of each microdialysis probe within the nucleus accumbens (all data shown in the same Paxinos atlas section; six were on the left side and one on the right). Abbreviations: DA, dopamine; 3-MT, 3-methoxytyramine; NE, norepinephrine; NM, normetanephrine; 5-HT, serotonin; DOPAC, 3,4-dihydroxyphenylacetate acid; HVA, homovanillic acid; 5HIAA, 5-hydroxyindole-3-acetic acid, GABA, ɣ-aminobutyric acid; ACh, acetylcholine. (b) Regression analysis results indicating strength of linear relationships between each analyte and each of four behavioral measures (reward rate; number of attempts; exploitation index; and cumulative rewards). Data are from 6 rats (7 sessions, total of 444 one-minute samples). Color scale shows p-values, Bonferroni-corrected for multiple comparisons (4 behavioral measures * 19 analytes), with red bars indicating a positive relationship and blue bars a negative relationship. Since both reward rate and attempts showed significant correlations with [DA], we constructed a regression model that included these predictors and an interaction term. In this model

R^2 remained at 0.15 and only reward rate showed a significant partial effect ($p < 2.38x10^{-12}$). (c) An alternative assessment of the relationship between minute-long [DA] samples and behavioral variables. Within each of the seven sessions [DA] levels were divided into three equal-sized bins (LOW, MEDIUM, HIGH); different colors indicate different sessions. For each behavioral variable, means were compared across [DA] levels using one-way ANOVA. There was a significant main effect of reward rate ($F(2,18)=10.02$, $p=0.0012$), but no effect of attempts ($F(2,18)=1.21$, $p=0.32$), exploitation index ($F(2,18)=0.081$, $p=0.92$), or cumulative rewards ($F(2,18)=0.181$, $p=0.84$). Post-hoc comparisons using the Tukey test revealed that the mean reward rates of LOW and HIGH [DA] differed significantly ($p=0.00082$). See also Supplementary Figs. 2,3.

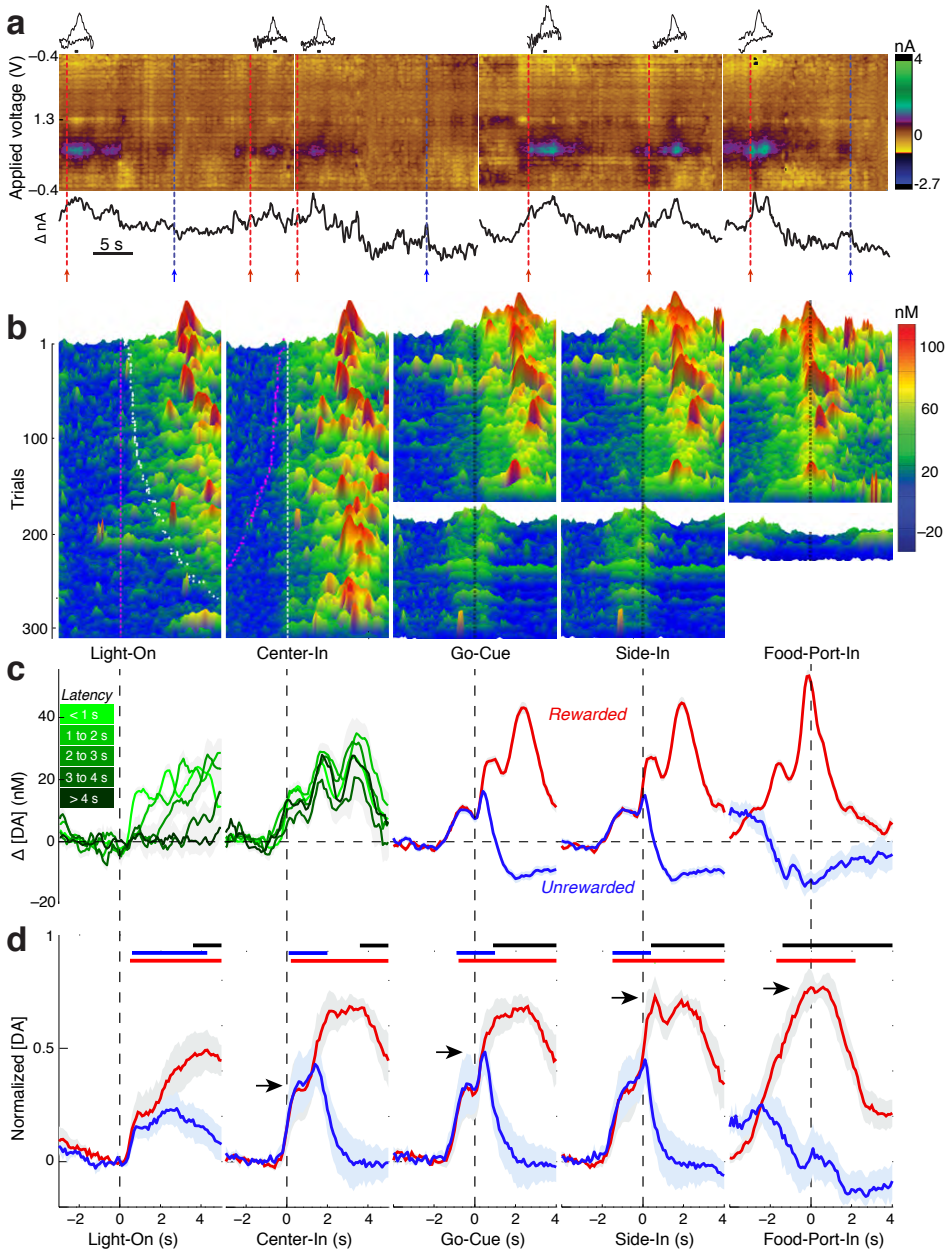

Figure 2.3: A succession of within-trial dopamine increases.

(a) Examples of FSCV data from a single session. Color plots display consecutive voltammograms (every 0.1s) as a vertical colored strip; examples of individual voltammograms are shown at top (taken from marked time points). Dashed vertical lines indicate Side-In events for rewarded (red) and unrewarded (blue) trials. Black traces

below indicate raw current values, at the applied voltage corresponding to the dopamine peak. (b) [DA] fluctuations for each of the 312 completed trials of the same session, aligned to key behavioral events. For Light-On and Center-In alignments, trials are sorted by latency (pink dots mark Light-On times; white dots mark Center-In times). For the other alignments rewarded (top) and unrewarded (bottom) trials are shown separately, but otherwise in the order in which they occurred. [DA] changes aligned to Light-On were assessed relative to a 2s baseline period, ending 1s before Light-On. For the other alignments, [DA] is shown relative to a 2s baseline ending 1s before Center-In. (c) Average [DA] changes during a single session (same data as b; shaded area represents s.e.m.). (d) Average event-aligned [DA] change across all six animals, for rewarded and unrewarded trials (see Figure 2.10 for each individual session). Data are normalized by the peak average rewarded [DA] in each session, and are shown relative to the same baseline epochs as in b. Black arrows indicate increasing levels of event-related [DA] during the progression through rewarded trials. Colored bars at top indicate time periods with statistically significant differences (red, rewarded trials greater than baseline, one-tailed t-tests for each 100ms time point individually; blue, same for unrewarded trials; black, rewarded trials different to unrewarded trials, 2-tailed t-tests; all statistical thresholds set to p=0.05, uncorrected).

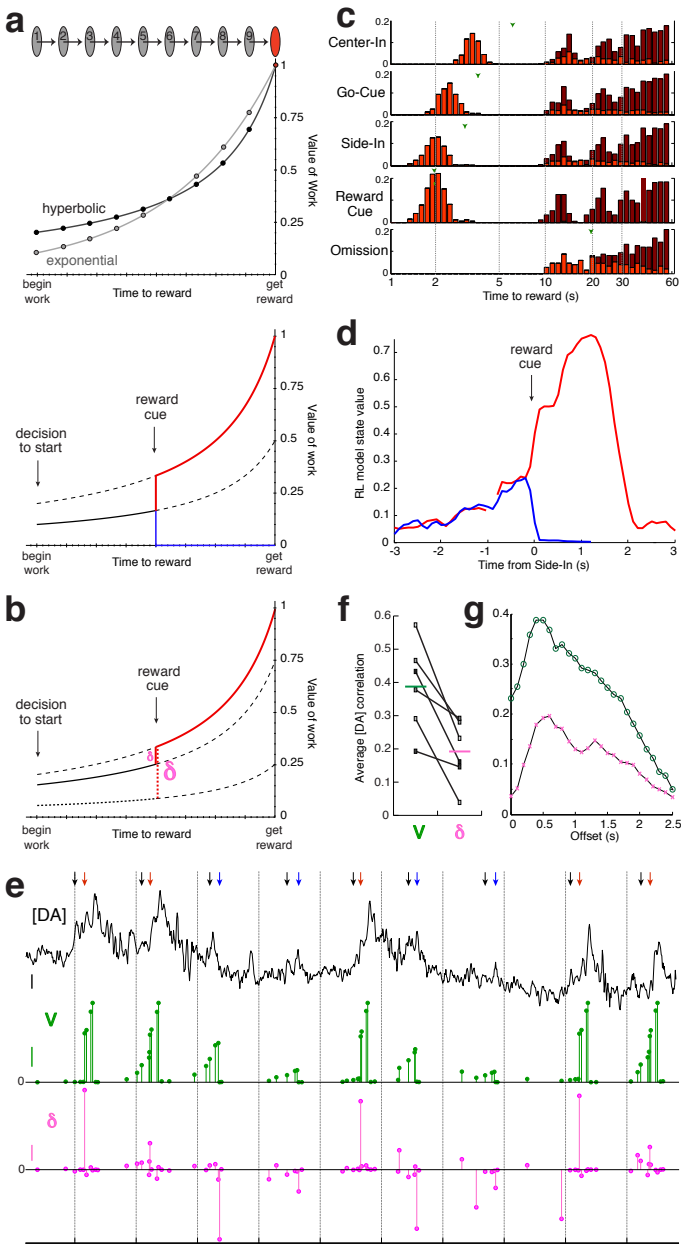

Figure 2.4: Within-trial dopamine fluctuations reflect state value dynamics.

(a) *Top,* Temporal discounting: the motivational value of rewards is lower when they are distant in time. With the exponential discounting commonly used in RL models, value is lower by a constant factor γ for each time step of separation from reward. People and other animals may actually use hyperbolic discounting which can optimize reward rate

(since rewards/time is inherently hyperbolic). Time parameters are here chosen simply to illustrate the distinct curve shapes. *Bottom.* Effect of reward cue, or omission, on state value. At trial start the discounted value of a future reward will be less if that reward is less likely. Lower value provides less motivational drive to start work - producing e.g. longer latencies. If a cue signals that upcoming reward is certain, the value function jumps up to the (discounted) value of that reward. For simplicity, the value of subsequent rewards is not included. (b) The reward prediction error δ reflects abrupt changes in state value. If the discounted value of work reflects an unlikely reward (e.g. probability = 0.25) a reward cue prompts a larger δ than if the reward was likely (e.g. probability = 0.75). Note that in this idealized example, δ would be zero at all other times. (c) *Top,* Task events signal updated times-to-reward. Data is from the same example session as Figure 2.3c. Bright red indicates times to the very next reward, dark red indicates subsequent rewards. Green arrowheads indicate average times to next reward (harmonic mean, only including rewards in the next 60s). As the trial progresses, average times-to-reward get shorter. If the reward cue is received, rewards are reliably obtained ~2s later. Task events are considered to prompt transitions between different internal states (Figure 2.11) whose learned values reflect these different experienced times-to-reward. (d) Average state value of the RL model for rewarded (red) and unrewarded (blue) trials, aligned on the Side-In event. The exponentially-discounting model received the same sequence of events as in Figure 2.3c, and model parameters (α=0.68, γ=0.98) were chosen for the strongest correlation to behavior (comparing state values at Center-In to latencies in this session, Spearman r=-0.34). Model values were binned at 100ms, and only bins with at least 3 events (state transitions) were plotted. (e) Example of the [DA] signal during a subset of trials from the same session, compared to model variables. Black arrows indicate Center-In events, red arrows Side-In with Reward Cue, blue arrows Side-In alone (Omission). Scale bars are: [DA], 20nM; V, 0.2; δ, 0.2. Dashed grey lines mark the passage of time in 10s intervals. (f) Within-trial [DA] fluctuations are more strongly correlated with model state value (V) than with RPE (δ). For every rat the [DA] : V correlation was significant (number of trials for each rat: 312, 229, 345, 252, 200, 204; $p<10^{-14}$ in each case; Wilcoxon signed-rank test of null hypothesis that median correlation within trials is zero) and significantly greater than the

[DA] : δ correlation ($p<10^{-24}$ in each case, Wilcoxon signed-rank test). Groupwise, both [DA] : V and [DA] : δ correlations were significantly non-zero, and the difference between them was also significant (n=6 sessions, all comparisons p=0.031, Wilcoxon signed-rank test). Model parameters (α=0.4, γ =0.95) were chosen to maximize the average behavioral correlation across all 6 rats (Spearman r = -0.28), but the stronger [DA] correlation to V than to δ was seen for all parameter combinations (Figure 2.11). (g) Model variables were maximally correlated with [DA] signals ~0.5s later, consistent with a slight delay caused by the time taken by the brain to process cues, and by the FSCV technique.

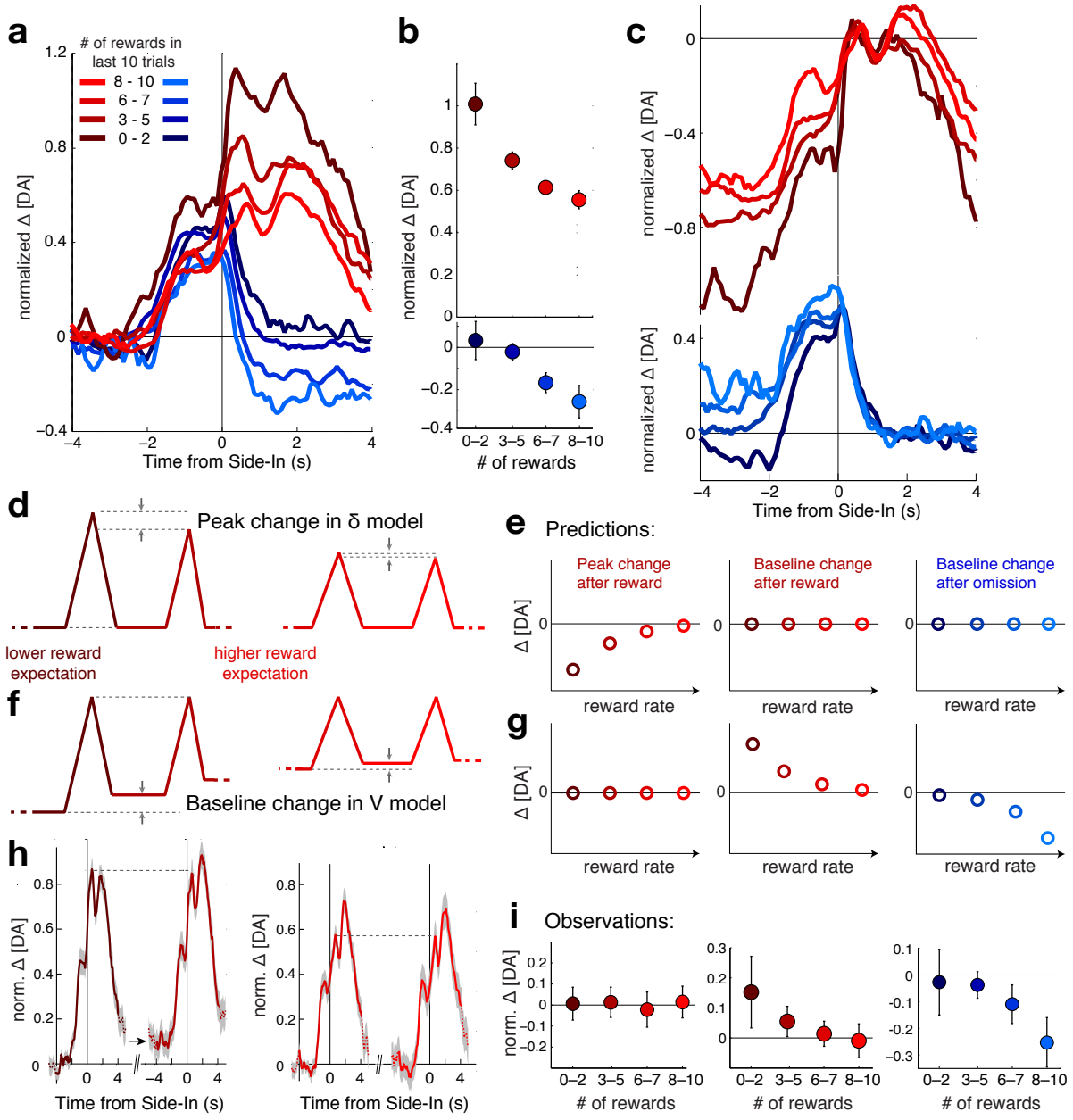

Figure 2.5: Between-trial dopamine shifts reflect updated state values.

(a) Less-expected outcomes provoke larger changes in [DA]. [DA] data from all FSCV sessions together (as in Figure 2.3d), broken down by recent reward history and shown relative to pre-trial “baseline" (-3 to -1s relative to Center-In). Note that the [DA] changes after reward omission last at least several seconds (shift in level), rather than showing a

highly transient dip followed by return to baseline as might be expected for encoding RPEs alone. (b) Quantification of [DA] changes, between baseline and reward feedback (0.5-1.0s after Side-In for rewarded trials, 1s-3s after Side-In for unrewarded trials). Error bars show SEM. (c) Same data as (a), but plotted relative to [DA] levels *after* reward feedback. These [DA] observations are consistent with a variable "baseline" whose level depends on recent reward history (as in Figure 2.4b model). (d) Alternative accounts of [DA] make different predictions for between-trial [DA] changes. When reward expectation is low, rewarded trials provoke large RPEs, but across repeated consecutive rewards RPEs should decline. Therefore if absolute [DA] levels encode RPE, the peak [DA] evoked by the reward-cue should decline between consecutive rewarded trials (and baseline levels should not change). For simplicity this cartoon omits detailed within-trial dynamics. (e) Predicted pattern of [DA] change under this account, which also does not predict any baseline shift after reward omissions (right). (f) If instead [DA] encodes state values, then peak [DA] should not decline from one reward to the next, but the baseline level should increase (and decrease following unrewarded trials). (g) Predicted pattern of [DA] change for this alternative account. (h) Unexpected rewards cause a shift in baseline, not in peak [DA]. Average FSCV data from consecutive pairs of rewarded trials (all FSCV sessions combined, as in a), shown relative to the pre-trial baseline of the first trial in each pair. Data were grouped into lower reward expectation (left pair of plots, 165 total trials; average time between Side-In events = 11.35s +/- 0.22s SEM) and higher reward expectation (right pair of plots, 152 total trials; time between Side-In events = 11.65s +/- 0.23s) by a median split of each individual session (using # rewards in last 10 trials). Dashed lines indicate that reward cues evoked a similar absolute level of [DA] in the second rewarded trial, compared to the first. Black arrow indicates the elevated pre-trial [DA] level for the second trial in the pair (mean change in baseline [DA] = 0.108, p=0.013, one-tailed Wilcoxon signed rank test). No comparable change was observed if the first reward was more expected (right pair of plots; mean change in baseline [DA] = 0.0013, p=0.108, one-tailed Wilcoxon signed rank test). (i) [DA] changes between consecutive trials follow the pattern expected for value coding, rather than RPE coding alone.

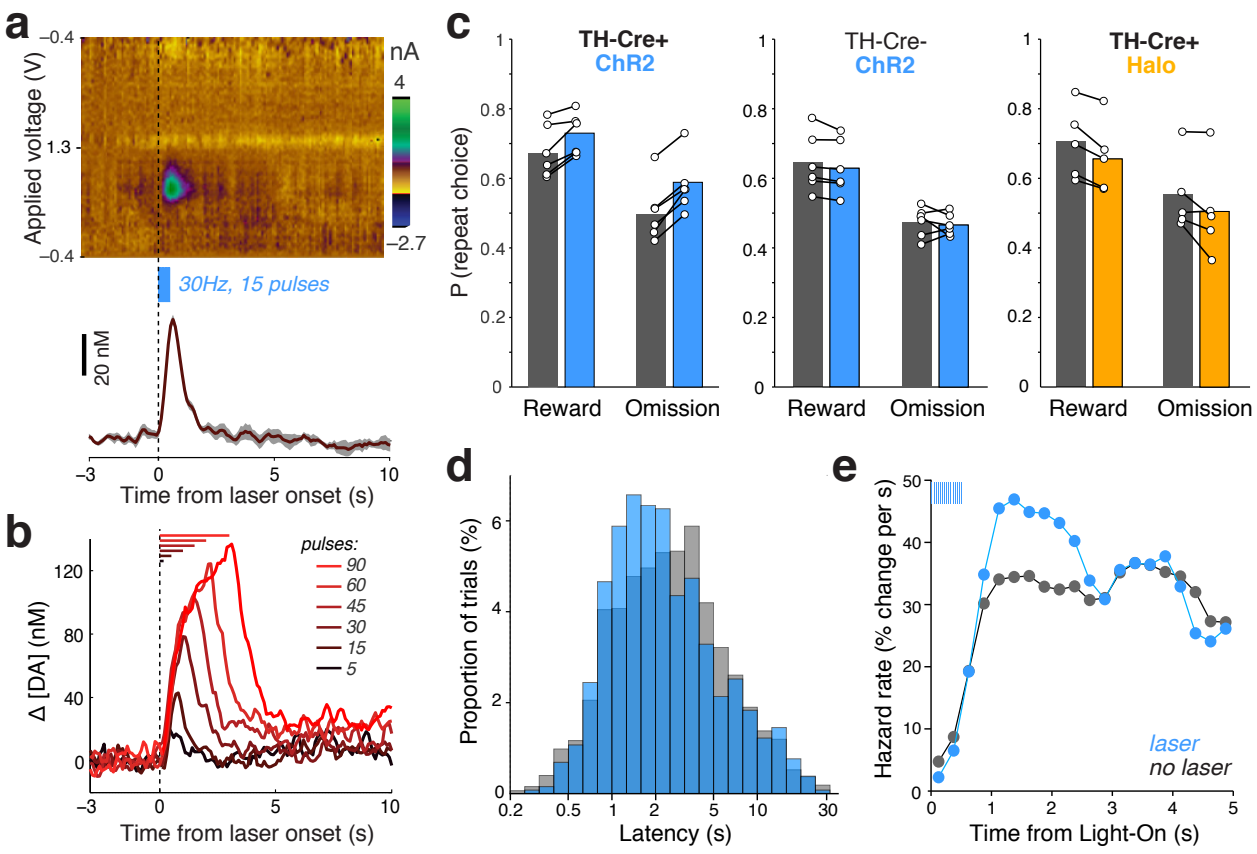

Figure 2.6: Phasic dopamine manipulations affect both learning and motivation.

(a) FSCV measurement of optogenetically-evoked [DA] increases. Optic fibers were placed above VTA, and [DA] change examined in nucleus accumbens core. Example shows dopamine release evoked by a 0.5s stimulation train (average of 6 stimulation events, shaded area indicates +/-SEM). (b) Effect of varying the number of laser pulses on evoked dopamine release, for the same 30Hz stimulation frequency. (c) Dopaminergic stimulation at Side-In reinforces the chosen left or right action. *Left,* in *TH-Cre*$^{+}$ rats stimulation of ChR2 increased the probability that the same action would be repeated on the next trial. Circles indicate average data for each of 6 rats (3 sessions each, 384 trials/session ± 9.5 SEM). *Middle,* this effect did not occur in *TH-Cre*$^{-}$ littermate controls (6 rats, 3 sessions each, 342±7 trials/session). *Right,* in *TH-Cre*$^{+}$ rats expressing Halorhodopsin, orange laser stimulation at Side-In reduced the chance

that the chosen action was repeated on the next trial (5 rats, 3 sessions each, 336±10 trials/session). See Figure 2.14 for additional analyses. (d) Laser stimulation at Light-On causes a shift towards sooner engagement, if the rats were not already engaged. Latency distribution (on log scale, 10 bins per log unit) for non-engaged, completed trials in *TH-Cre*$^{+}$ rats with ChR2 (n=4 rats with video analysis; see Figure 2.15 for additional analyses). (e) Same latency data as d, but presented as hazard rates. Laser stimulation (blue ticks at top left) increases the chance that rats will decide to initiate an approach, resulting in more Center-In events 1-2s later (for these n=4 rats, one-way ANOVA on hazard rate $F(1,3) = 18.1$, $p=0.024$). See Figure 2.70 for hazard rate time courses from the individual rats.

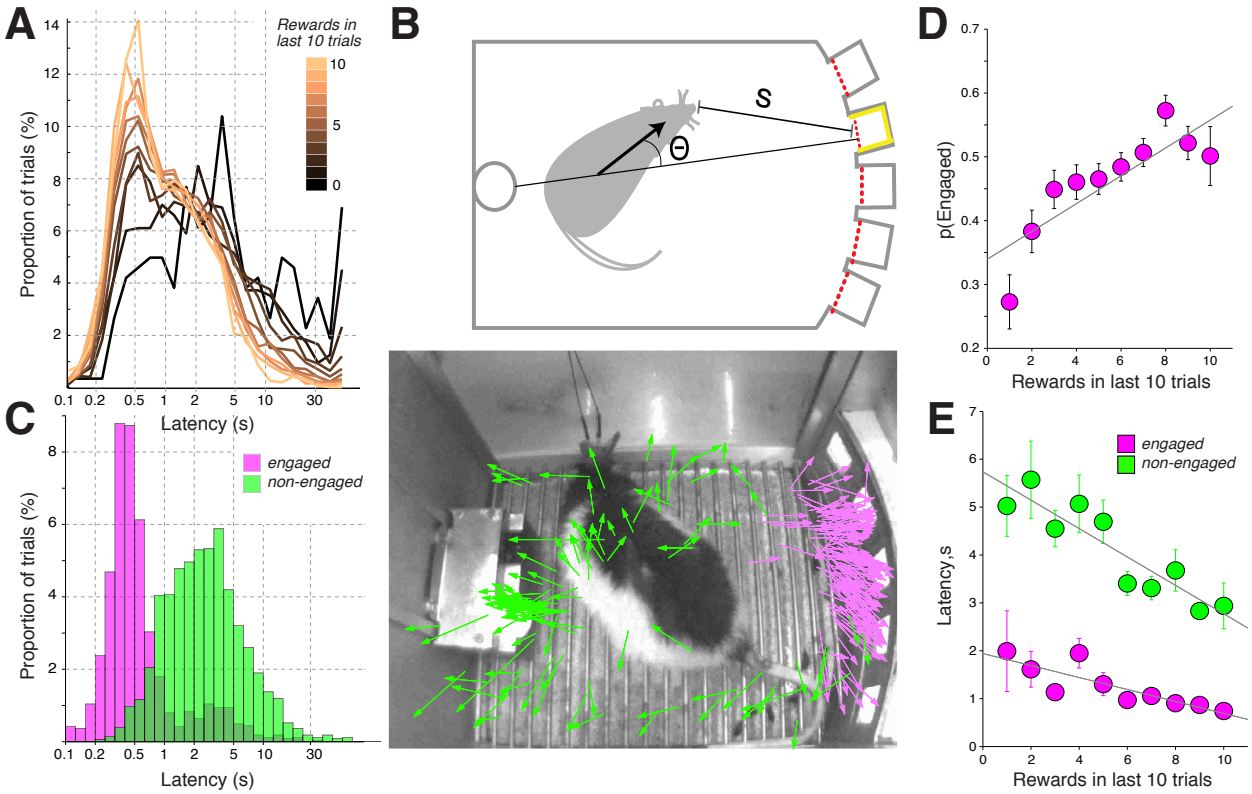

Figure 2.7: Reward rate affects the decision to begin work.

(a) Latency distributions are bimodal, and depend on reward rate. Very short latencies (early peak) preferentially occur when a greater proportion of recent trials have been rewarded (same data set as Fig 1**d-i**). (**b**) (top) Schematic of video analysis. Each trial was categorized as "engaged" (already waiting for Light-On) or non-engaged based upon distance (s) and orientation (θ) immediately before Light-On (see Methods). (bottom) Arrows indicate rat head position and orientation for engaged (pink) and non-engaged (green) trials (one example session shown). (**c**) Categorization into engaged, non-engaged trials accounts for bimodal latency distribution (data shown are all non-laser trials across 12 ChR2 sessions in *TH-Cre+* rats). (**d**) Proportion of engaged trials increases when more recent trials have been rewarded (3336 trials from 4 rats, r=0.82,

p=0.003). (**e**) Especially for non-engaged trials, latencies are lower when reward rate is higher (r=-0.11,p=0.004 for 1570 engaged trials, r=-0.18, $p=5.2x10^{-19}$ for 1766 non-engaged trials).

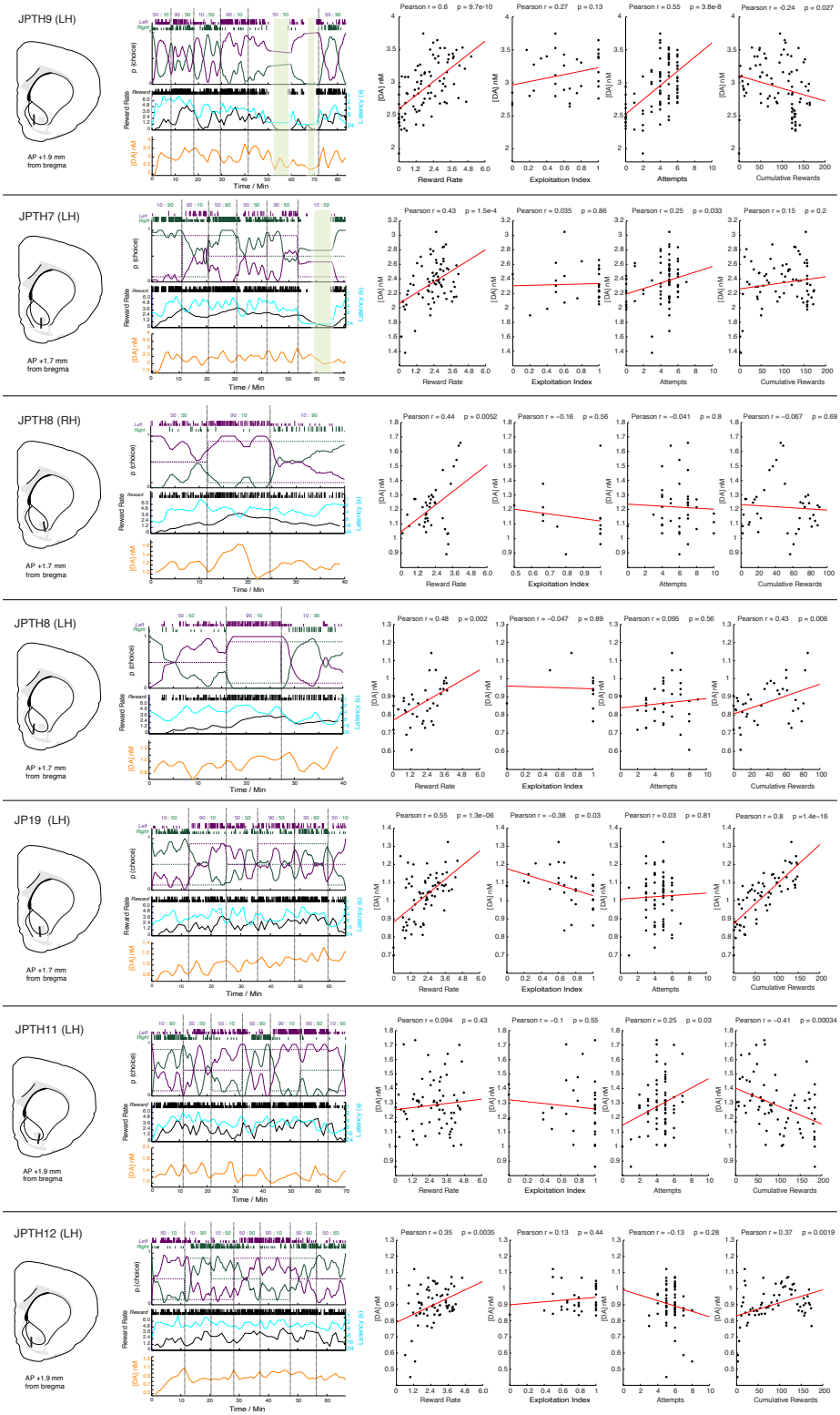

Figure 2.8: Individual microdialysis sessions.

Each row shows data for a different session, with indicated rat ID (e.g. IM463) and recording side (LH = left, RH=right). From left: dialysis probe location, behavioral and [DA] time courses, and individual session correlations to behavioral variables. Reward rate is in units of rewards per min. Numbers of microdialysis samples for each of the seven sessions: 86,72,39,39,68,73,67 respectively. The overall relationship between dopamine and reward rate remained highly significant even if excluding periods of inactivity (defined as no trials initiated for >2 minutes, shaded in green; regression R^2 = 0.12, $p = 1.4 \times 10^{-13}$).

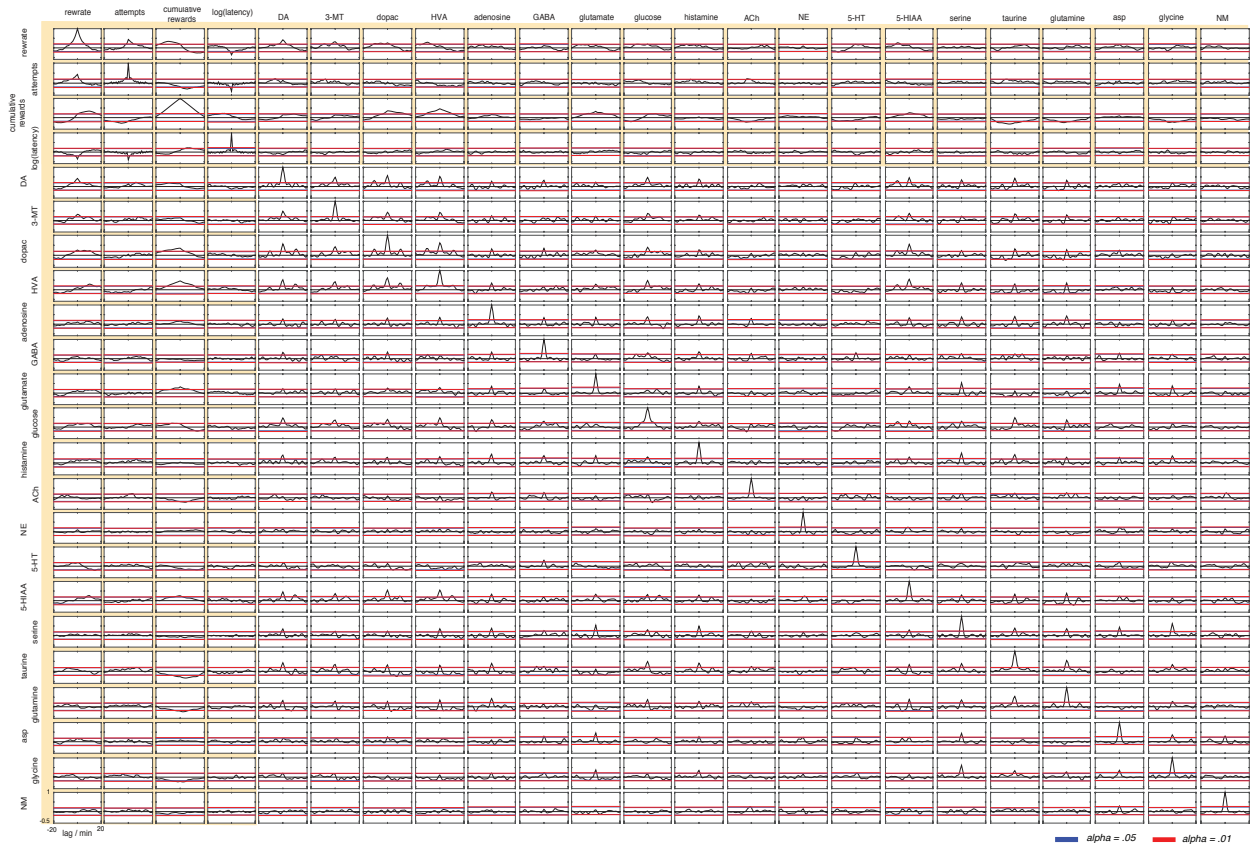

Figure 2.9: Cross-correlograms for behavioral variables and neurochemicals.

Each plot shows cross-correlograms averaged across all microdialysis sessions, all using the same axes (-20min to +20min lags, -0.5 to +1 correlation). Colored lines indicate statistical thresholds corrected for multiple comparisons (see Methods). Many neurochemical pairs show no evidence of covariation, but others display strong relationships including a cluster of glutamate, serine, aspartate and glycine.

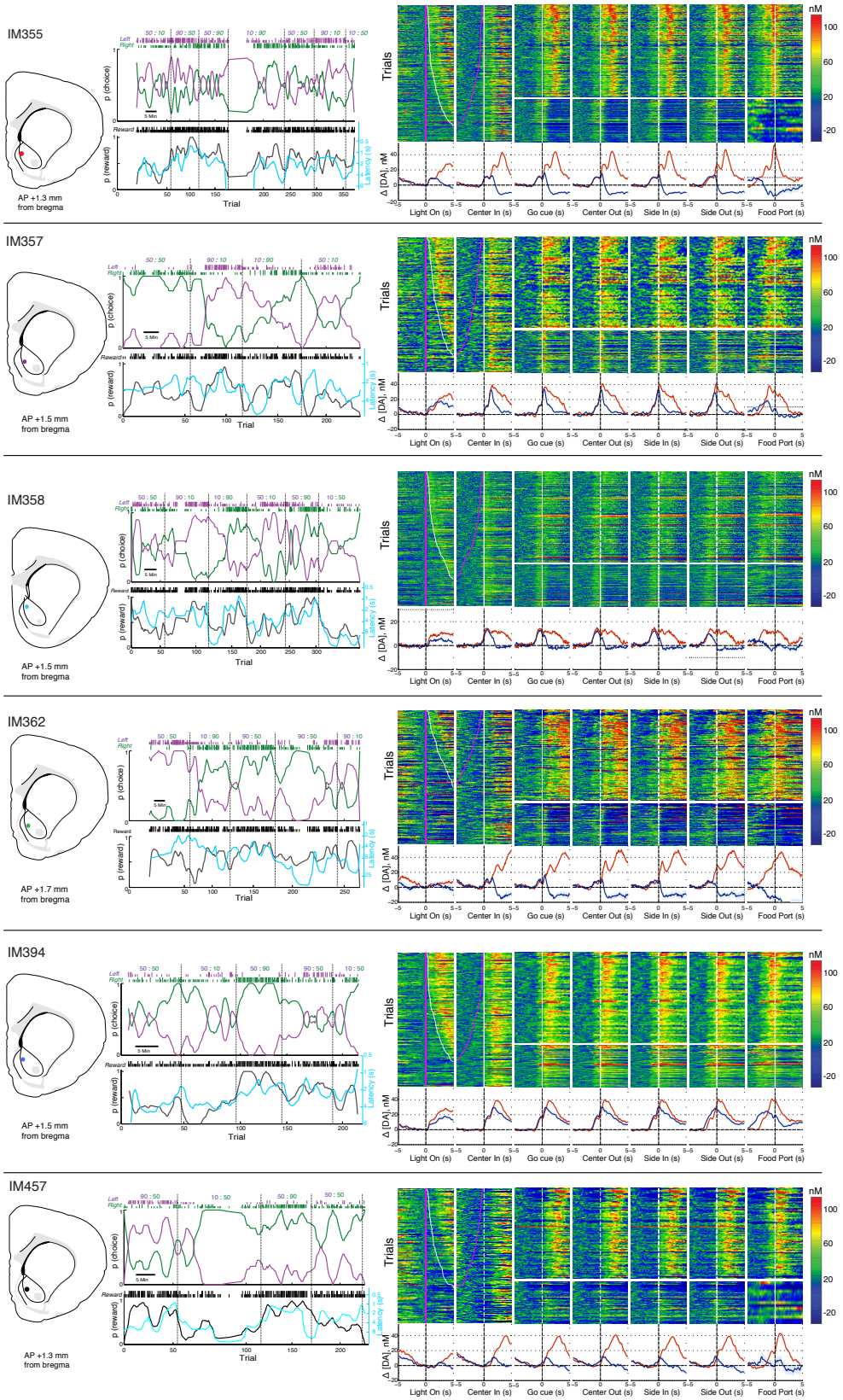

Figure 2.10: Individual voltammetry sessions.

Each row shows data for a different rat (e.g. IM355, which was also used as the example in Figs.3,4). At left, recording site within nucleus accumbens. Middle panels show behavioral data for the FSCV session (same format as Figure 2.1). Right panels show individual FSCV data (same format as Figure 2.3, but with additional event alignments).

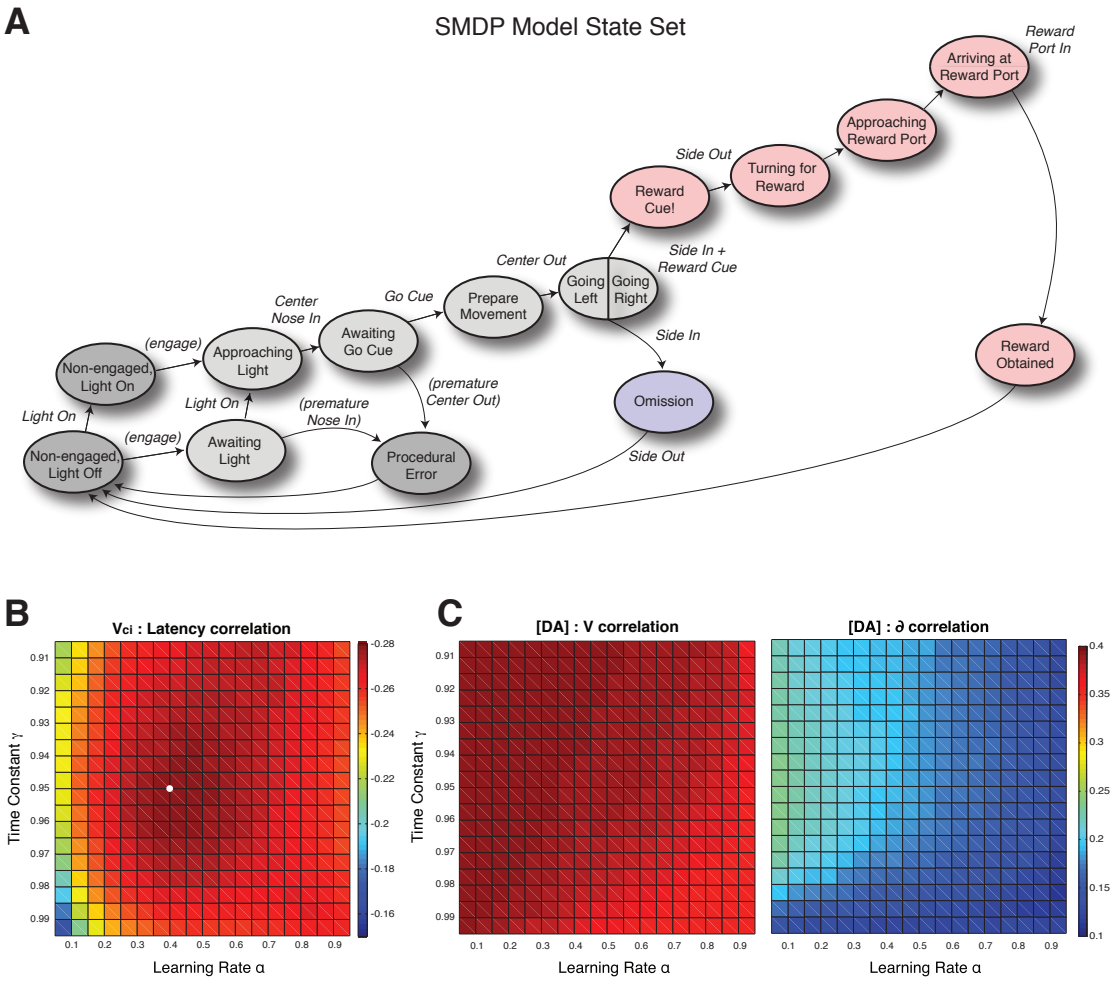

Figure 2.11: SMDP model.

(**a**) Task performance was modeled as a sequence of transitions between states of the agent (rat). Each state had a single associated cached value V(s) (rather than, for example, separate state-action (Q) values for leftward and rightward trials). Most state transitions occur at variable times (hence “semi-Markov”) marked by observed external events (Center-In, Go-Cue, etc). In contrast, the state sequence between Side-Out and Reward Port In is arbitrarily defined (“Approaching Reward Port” begins 1s before Reward Port In; “Arriving At Reward Port” begins 0.5s before Reward Port In). Changing

the number or specific timing of these intermediate states does not materially affect the rising shape of the value function. (**b**) Average correlation (color scale = Spearman's r) between SMDP model state value at Center-In (Vci) and latency across all six FSCV rats, for a range of learning rates α and exponential discounting time constants ɣ. Note that color scale is inverted (red indicates strongest negative relationship, with higher value corresponding to shorter latency). White dot marks point of strongest relationship (α=0.40, ɣ=0.95). (**c**) Correlation between [DA] and state value V is stronger than the correlation between [DA] and reward prediction error δ, across the same range of parameters. Color scale at right is the same for both matrices (Spearman's r).

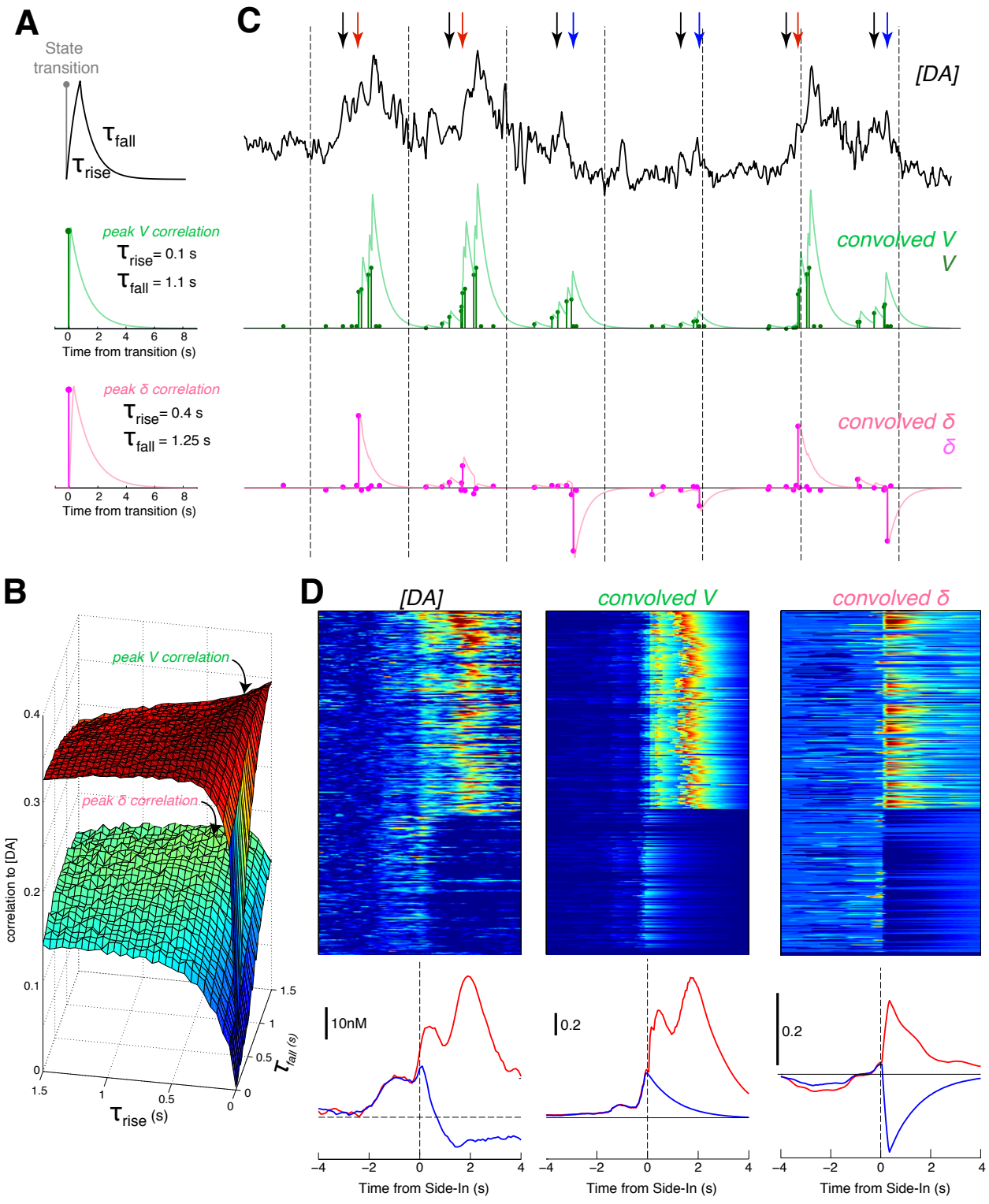
A
State transition
τfall
τrise
peak V correlation
τrise= 0.1 s
τfall = 1.1 s
0 2 4 6 8
Time from transition (s)
peak δ correlation
τrise= 0.4 s
τfall = 1.25 s
0 2 4 6 8
Time from transition (s)
C
[DA]
convolved V
V
convolved δ
δ
B
peak V correlation
peak δ correlation
correlation to [DA]
0.4
0.3
0.2
0.1
0
1.5
1
0.5
0
τrise (s)
τfall (s)
D
[DA]
convolved V
convolved δ
10nM
0.2
0.2
−4 −2 0 2 4
Time from Side-In (s)
−4 −2 0 2 4
Time from Side-In (s)
−4 −2 0 2 4
Time from Side-In (s)

Figure 2.12: Dopamine relationship to temporals-stretched model variables:

*(***a**) Kernel consisted of an exponential rise (to 50% of asymptote) and an exponential fall, with separate time constants. (**b**) Within-trial correlation coefficients between [DA] and kernel-convolved model variables V and δ, for a range of rise and fall time constants (0 - 1.5s each, in 50ms timesteps, using data from all 6 rats). Regardless of parameter values, [DA] correlations to V were always higher than to δ. (**c**) Same example data as Fig. 4E, but also showing convolved V and δ (using time constants that maximized correlation to [DA] in each case). (**d**) Trial-by-trial (top) and average (bottom) [DA], convolved V, and convolved δ, for the same session as Fig. 4**d**,**e**.

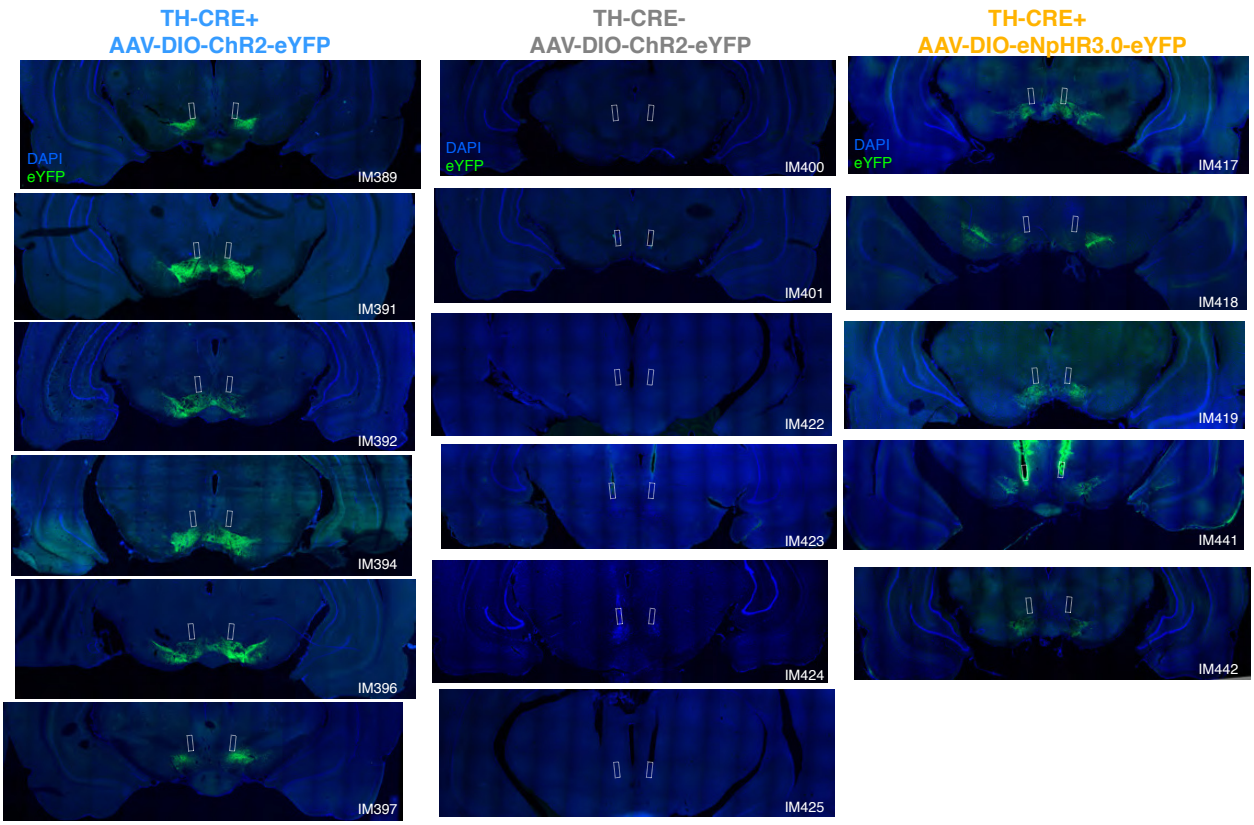

Figure 2.13: Histology for behavioral ontogenetic experiments.

Identifier (e.g. “IM389”) for each rat is given at bottom right corner. Coronal sections shown are within 180μm (anterior-posterior) of the observed fiber tip location. Green indicates expression of eYFP, blue is DAPI counterstain. In a couple of cases (IM423, IM441) autofluorescence of damaged brain tissue is visible along the optic fiber tracts; this was not specific to the green channel.

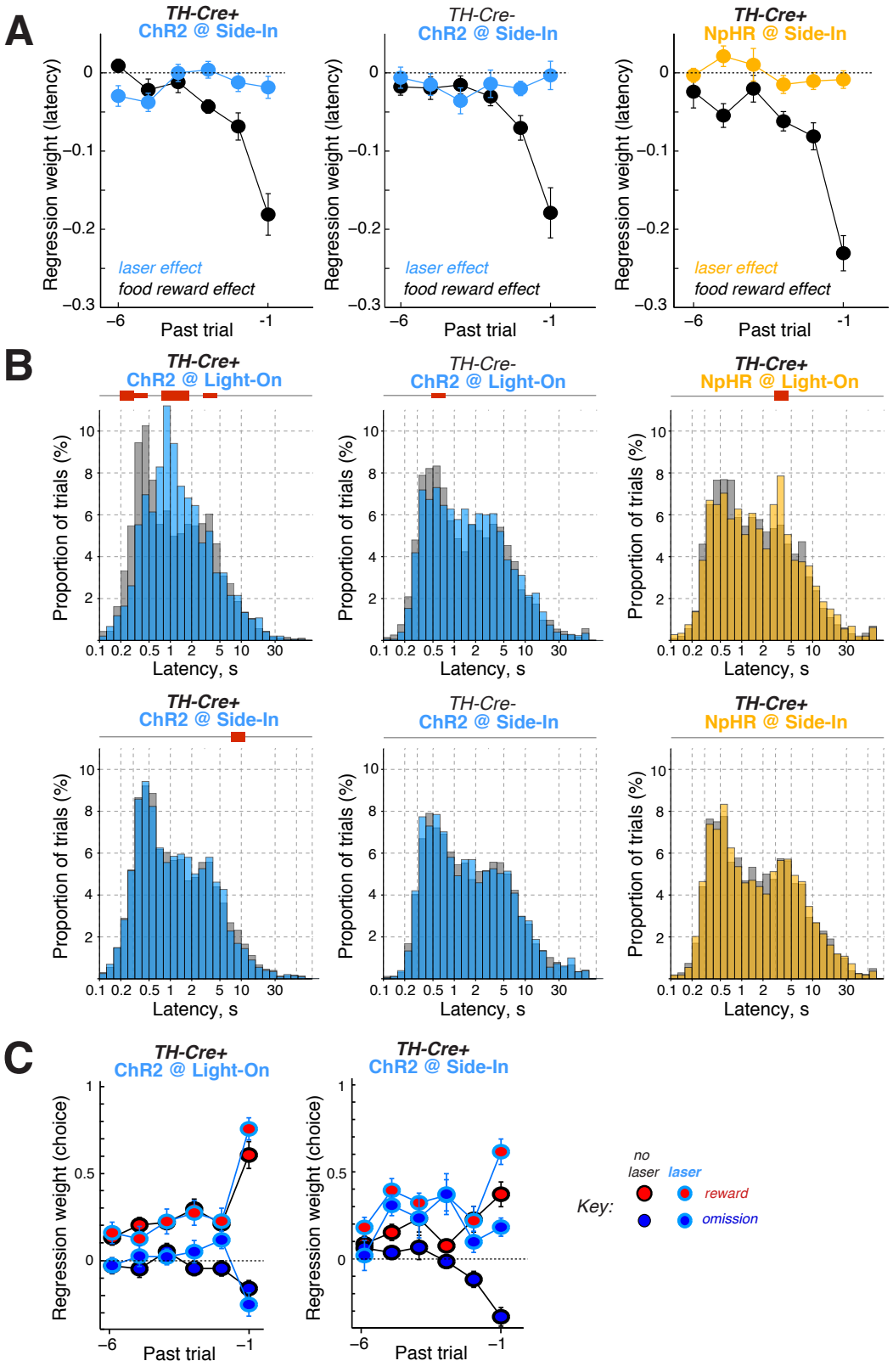

Figure 2.14: Further analysis of persistence of optogenetic effects.

(**a**) Regression analysis showing substantial effects of recent rewards (black) on latency, but no comparable effect of recent Side-In laser stimulations on latency. (**b**) Effects of Light-On [DA] manipulation on same-trial latency distributions (top), and of Side-In [DA] manipulation on next-trial latency distributions (bottom). Dataset shown is the same as Fig. 6c, i.e. all completed trials in TH-Cre^{+} rats with ChR2 (left), TH-Cre^{-} rats with ChR2 (middle) and TH-Cre^{+} rats with halorhodopsin (right). (**c**) Regression analysis of laser stimulation on subsequent left/right choices. Recent food rewards for a given left/right action increase the probability that it will be repeated. Extra [DA] at Light-On has little or no effect on subsequent choices, but extra [DA] at Side-In is persistently reinforcing. For the Side-In data, note especially the positive coefficients for otherwise unrewarded laser trials.

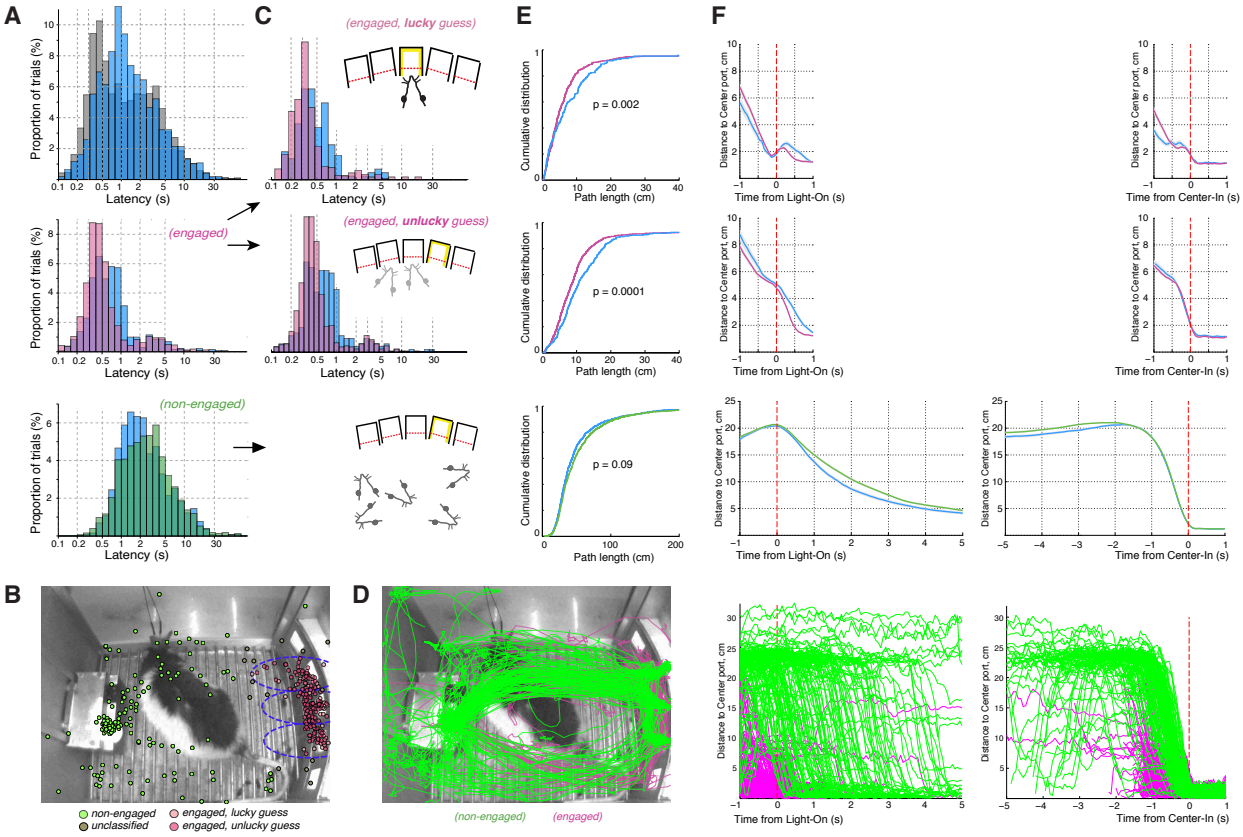

Figure 2.15: Video analysis of ontogenetic effects on latency.

(**a**) Extra [DA] at Light-On causes shorter latencies for non-engaged trials, but longer latencies for a subset of engaged trials. Top plot shows all trials (for the n=4 *TH-Cre*+ rats with ChR2 stimulation at Light-On for which video was recorded; 3 sessions/rat; 3336 no-laser trials in grey; 1335 laser trials in blue). Bottom plots show the breakdown into engaged (n=1975) and non-engaged (n=2696) trials. (**b**) We examined whether laser-slowed trials might be those in which the rat was waiting at the wrong port (if, for example, DA were to increase the salience of currently attended stimuli). Engaged trials were further broken down into “lucky guesses” (those trials for which the rat was immediately adjacent to the start port as it was illuminated) and “unlucky guesses” (immediately adjacent to one of the other two possible start ports). Blue dashed ellipses indicate zones used to classify trials by guessed port (8.5cm long

diameter, 3.4cm short diameter) (**c**) Laser-slowing was observed for both lucky (n=603) and unlucky (n=1007) guesses. Note that blue distribution is bimodal in both cases, indicating that only a subset of trials were affected. Video observations suggested that on some trials extra [DA] evokes a small extra head/neck movement, that makes the trajectory to the illuminated port longer and therefore slower. (**d**) Quantification of trajectories, by scoring rat location on each video frame from 1s before Light-On to 1s after Center-In. Colored lines show all individual trajectories for one example session. Panels at right show the same trajectories plotted as distance remaining from Center-In port, by time elapsed from either Light-On or Center-In. Note that for non-engaged trials (green), the approach to the Center-In port consistently takes ~1-2s. Therefore, the epoch considered as "baseline" in the FSCV analyses (-3 to -1s relative to Center-In) is around the time that rats decide to initiate approach behaviors. (**e**) Extra [DA] causes longer average trajectories for engaged trials. Cumulative distributions of path-lengths between Light-On and Center-In, for (top-to-bottom) engaged/lucky, engaged/unlucky and non-engaged respectively. Blue lines indicate laser trials, and p-values are from Komolgorov-Smirnov tests comparing laser to no-laser distributions (no-laser/laser trial numbers: top, 292/75; middle, 424/99; bottom, 1897/792). On engaged trials rats often reoriented between the three potential start ports, perhaps checking if they were illuminated; one possibility is that the extra laser-evoked movement on engaged trials reflects dopaminergic facilitation of these orienting movements. If such a movement is already close to execution before Light-On, it may be evoked before the correct start port can be appropriately targeted. (**f**) Additional trajectory analysis, plotting time courses of rat distance from the illuminated start port. On non-engaged trials extra [DA] tends to make the approach to the illuminated start port occur earlier (note progressive separation of green, blue lines when aligned on Light-On). However, the approach time course is extremely similar (note overlapping lines in the final ~1-2s before Center-In), indicating that extra [DA] did not affect the speed of approach.

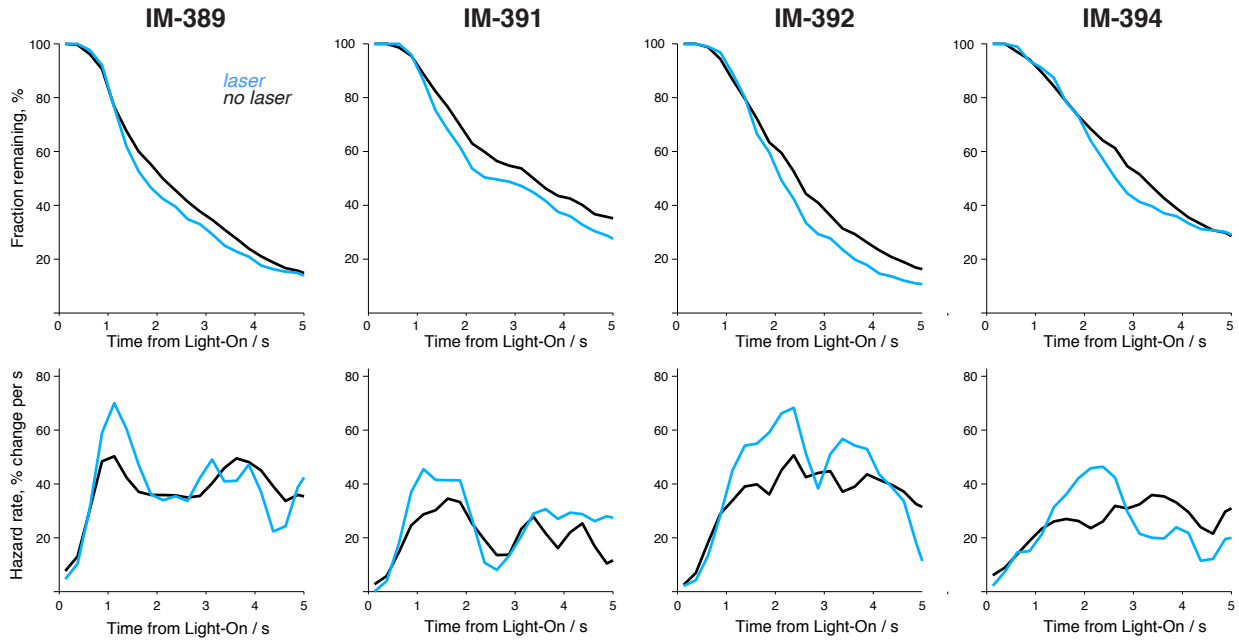

Figure 2.16: Optogenetic effects on hazard rates for individual video-scored rats.

Latency survivor plots (top) and corresponding hazard rates (bottom) for each of the four *TH-Cre+* rats with ChR2 stimulation at Light-On for which video was recorded (each rat had 3 video sessions that were concatenated for analysis). Only non-engaged trials are included (Numbers of no-laser/laser trials: IM-389, 522/215; IM-391, 294/125; IM-392, 481/191; IM-394, 462/189). For each rat laser stimulation caused an increase in the hazard rate of the Center-In event ~1-2s later (the duration of an approach).

References

Adamantidis, A.R., Tsai, H.C., Boutrel, B., Zhang, F., Stuber, G.D., Budygin, E.A., Touriño, C., Bonci, A., Deisseroth, K., and de Lecea, L. (2011). Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. J Neurosci *31*, 10829-10835.

Ainslie, G. (2005). Précis of breakdown of will. Behavioral and Brain Sciences *28*, 635-650.

Aragona, B.J., Day, J.J., Roitman, M.F., Cleaveland, N.A., Wightman, R.M., and Carelli, R.M. (2009). Regional specificity in the real-time development of phasic dopamine transmission patterns during acquisition of a cue-cocaine association in rats. Eur J Neurosci *30*, 1889-1899.

Beeler, J.A., Frazier, C.R., and Zhuang, X. (2012). Putting desire on a budget: dopamine and energy expenditure, reconciling reward and resources. Frontiers in integrative neuroscience *6*, 49.

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., and Dayan, P. (2013). Dopamine modulates reward-related vigor. Neuropsychopharmacology *38*, 1495-1503.

Berridge, K.C. (2007). The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) *191*, 391-431.

Cagniard, B., Balsam, P.D., Brunner, D., and Zhuang, X. (2006). Mice with chronically elevated dopamine exhibit enhanced motivation, but not learning, for a food reward. Neuropsychopharmacology *31*, 1362-1370.

Cannon, C.M., and Palmiter, R.D. (2003). Reward without dopamine. The Journal of neuroscience *23*, 10827-10831.

Daw, N.D., Courville, A.C., Tourtezky, D.S., and Touretzky, D.S. (2006). Representation and timing in theories of the dopamine system. Neural Comput *18*, 1637-1677.

Daw, N.D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural Netw *15*, 603-616.

Day, J.J., Roitman, M.F., Wightman, R.M., and Carelli, R.M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci *10*, 1020-1028.

Dreyer, J.K., Herrik, K.F., Berg, R.W., and Hounsgaard, J.D. (2010). Influence of phasic and tonic dopamine release on receptor activation. The Journal of Neuroscience *30*, 14273-14283.

Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898-1902.

Freed, C.R., and Yamamoto, B.K. (1985). Regional brain dopamine metabolism: a marker for the speed, direction, and posture of moving animals. Science *229*, 62-65.

Gage, G.J., Stoetzner, C.R., Wiltschko, A.B., and Berke, J.D. (2010). Selective activation of striatal fast-spiking interneurons during choice execution. Neuron *67*, 466-479.

Gan, J.O., Walton, M.E., and Phillips, P.E. (2010). Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. Nat Neurosci *13*, 25-27.

Gershman, S.J. (2014). Dopamine ramps are a consequence of reward prediction errors. Neural Comput *26*, 467-471.

Guitart-Masip, M., Beierholm, U.R., Dolan, R., Duzel, E., and Dayan, P. (2011). Vigor in the face of fluctuating rates of reward: an experimental examination. Journal of cognitive neuroscience *23*, 3933-3938.

Haith, A.M., Reppert, T.R., and Shadmehr, R. (2012). Evidence for hyperbolic temporal discounting of reward in control of movements. The Journal of Neuroscience *32*, 11727-11736.

Hart, A.S., Rutledge, R.B., Glimcher, P.W., and Phillips, P.E. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci *34*, 698-704.

Heien, M.L., Khan, A.S., Ariansen, J.L., Cheer, J.F., Phillips, P.E.M., Wassum, K.M., and Wightman, R.M. (2005). Real-time measurement of dopamine fluctuations after cocaine in the brain of behaving rats. Proc Natl Acad Sci U S A *102*, 10023-10028.

Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E., and Graybiel, A.M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature

Hull, C.L. (1932). The goal-gradient hypothesis and maze learning. Psychological Review *39*, 25.

Humphries, M.D., Khamassi, M., and Gurney, K. (2012). Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. Front Neurosci *6*, 9.

Ishiwari, K., Weber, S.M., Mingote, S., Correa, M., and Salamone, J.D. (2004). Accumbens dopamine and the regulation of effort in food-seeking behavior: modulation of work output by different ratio or force requirements. Behav Brain Res *151*, 83-91.

Kable, J.W., and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. Nat Neurosci *10*, 1625-1633.

Kacelnik, A. (1997). Normative and descriptive models of decision making: time discounting and risk sensitivity. Characterizing human psychological adaptations *208*, 51-66.

Kile, B.M., Walsh, P.L., McElligott, Z.A., Bucher, E.S., Guillot, T.S., Salahpour, A., Caron, M.G., and Wightman, R.M. (2012). Optimizing the temporal resolution of fast-scan cyclic voltammetry. ACS chemical neuroscience *3*, 285-292.

Kim, K.M., Baratta, M.V., Yang, A., Lee, D., Boyden, E.S., and Fiorillo, C.D. (2012). Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PLoS One *7*, e33612.

Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. J Neurosci *28*, 7837-7846.

Lau, B., and Glimcher, P.W. (2005). Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys. Journal of the Experimental Analysis of Behavior *84*, 555-579.

Leventhal, D.K., Gage, G.J., Schmidt, R., Pettibone, J.R., Case, A.C., and Berke, J.D. (2012). Basal ganglia beta oscillations accompany cue utilization. Neuron *73*, 523-536.

Leventhal, D.K., Stoetzner, C.R., Abraham, R., Pettibone, J., DeMarco, K., and Berke, J.D. (2014). Dissociable effects of dopamine on learning and performance within sensorimotor striatum. Basal Ganglia *4*, 43-54.

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837-841.

Mazur, J.E. (1984). Tests of an equivalence rule for fixed and variable reinforcer delays. Journal of Experimental Psychology: Animal Behavior Processes *10*, 426.

McClure, S.M., Daw, N.D., and Read Montague, P. (2003). A computational substrate for incentive salience. Trends Neurosci *26*, 423-428.

Morita, K., and Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits *8*, 36.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat Neurosci *9*, 1057-1063.

Nagano-Saito, A., Cisek, P., Perna, A.S., Shirdel, F.Z., Benkelfat, C., Leyton, M., and Dagher, A. (2012). From anticipation to action, the role of dopamine in perceptual decision making: an fMRI-tyrosine depletion study. J Neurophysiol *108*, 501-512.

Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl) *191*, 507-520.

Niyogi, R.K., Breton, Y.-A., Solomon, R.B., Conover, K., Shizgal, P., and Dayan, P. (2014). Optimal indolence: a normative microscopic approach to work and leisure. Journal of The Royal Society Interface *11*, 20130969.

Phillips, P.E., Stuber, G.D., Heien, M.L., Wightman, R.M., and Carelli, R.M. (2003). Subsecond dopamine release promotes cocaine seeking. Nature *422*, 614-618.

Rapoport, J.L., Buchsbaum, M.S., Weingartner, H., Zahn, T.P., Ludlow, C., and Mikkelsen, E.J. (1980). Dextroamphetamine: Its cognitive and behavioral effects in normal and hyperactive boys and normal men. Archives of General Psychiatry *37*, 933-943.

Reynolds, J.N., Hyland, B.I., and Wickens, J.R. (2001). A cellular mechanism of reward-related learning. Nature *413*, 67-70.

Roitman, M.F., Stuber, G.D., Phillips, P.E., Wightman, R.M., and Carelli, R.M. (2004). Dopamine operates as a subsecond modulator of food seeking. J Neurosci *24*, 1265-1271.

Salamone, J.D., and Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. Neuron *76*, 470-485.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. Science *310*, 1337-1340.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci *23*, 9913-9923.

Schlinger, H.D., Derenne, A., and Baron, A. (2008). What 50 years of research tell us about pausing under ratio schedules of reinforcement. Behavior Analyst *31*, 39.

Schmidt, R., Leventhal, D.K., Mallet, N., Chen, F., and Berke, J.D. (2013). Canceling actions involves a race between basal ganglia pathways. Nat Neurosci *16*, 1118-1124.

Schultz, W. (1997). A Neural Substrate of Prediction and Reward. Science *275*, 1593-1599.

Simen, P., Cohen, J.D., and Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. Neural Netw *19*, 1013-1026.

Song, P., Mabrouk, O.S., Hershey, N.D., and Kennedy, R.T. (2011). In vivo neurochemical monitoring using benzoyl chloride derivatization and liquid chromatography--mass spectrometry. Analytical chemistry *84*, 412-419.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci *16*, 966-973.

Stephens, D.W., and Krebs, J.R. (1986). Foraging theory (Princeton University Press).

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. Science *304*, 1782-1787.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning : an introduction (Cambridge, Mass.: MIT Press).

Tachibana, Y., and Hikosaka, O. (2012). The primate ventral pallidum encodes expected reward value and regulates motor action. Neuron *76*, 826-837.

Threlfell, S., Lalic, T., Platt, N.J., Jennings, K.A., Deisseroth, K., and Cragg, S.J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron *75*, 58-64.

van Der Meer, M.A., and Redish, A.D. (2011). Ventral striatum: a critical look at models of learning and evaluation. Current opinion in neurobiology *21*, 387-392.

Venton, B.J., Troyer, K.P., and Wightman, R.M. (2002). Response times of carbon fiber microelectrodes to dynamic changes in catecholamine concentration. Analytical chemistry *74*, 539-546.

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

Wardle, M.C., Treadway, M.T., Mayo, L.M., Zald, D.H., and de Wit, H. (2011). Amping up effort: effects of d-amphetamine on human effort-based decision-making. The Journal of Neuroscience *31*, 16597-16602.

Witten, I.B., Steinberg, E.E., Lee, S.Y., Davidson, T.J., Zalocusky, K.A., Brodsky, M., Yizhar, O., Cho, S.L., Gong, S., Ramakrishnan, C., Stuber, G.D., Tye, K.M., Janak, P.H., and Deisseroth, K. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron *72*, 721-733.

CHAPTER 3: PLASTICITY WINDOWS FOR DOPAMINE REINFORCEMENT OF STATE AND ACTION VALUES

Introduction

Evaluative-feedback recruits many brain circuits involved in valuation mechanisms used for adjusting performance vigor and adapting choice behaviors. Thus, actions followed by rewards are more likely to occur in the future (i.e reinforcement learning), and dopamine (DA) is critical for learning from rewards. The precise contributions of mesolimbic DA in motivation, associative leaning and reinforcement continue to be debated (Berridge, 2007; Schultz, 2007; Salamone and Correa, 2012), but DA is causal to key aspects of dissociable learning and motivational processes.

In chapter two, I described results from several experiments using a range of techniques directed at assessing the apparent simultaneous (and complementary) contributions of mesolimbic DA to learning and motivation. We reported that, across multiple timescales, DA signals a single decision variable: a Value function (an ongoing estimate of discounted future rewards) that is used for motivational decision making ('Is It worth it?') and abrupt changes in this value function serve as temporal-difference reward prediction errors used for learning ("repeat action?").

We measured sub-second accumbens DA fluctuations and found that pre-trial levels covary with overall reward expectation (state-value), and predict approach behaviors. Furthermore, brief optogenetic stimulation (or inhibition) of VTA DA cells (bidirectionally) affected rats' immediate willingness to engage in task performance as indexed by changes in latency to initiate the current trial. This indicated that early trial

DA is both necessary and sufficient to influence instantaneous behavioral expressions of learned state-values.

During the feedback epoch, we found that more surprising outcomes caused larger increases in striatal DA fluctuations, consistent with coding of reward prediction error (RPE) signals. These outcome-related increases in DA were correlated to the extent of choice reinforcement on the next trial (i.e increased action-values). Furthermore, artificially increasing (or decreasing) outcome related DA changes via optogenetic manipulations (bidirectionally) affected the likelihood of choice re-section on the next trial. Together these results support the notion that reward-feedback related DA is both necessary and sufficient for shaping future decisions by updating the stored value of executed actions (Maia and Frank, 2011; Adamantidis et al., 2011; Kim et al., 2012; Steinberg et al., 2013; Witten et al., 2011; Tsai et al., 2009).

While our results provide evidence for how DA fluctuations exert activational (by promoting immediate motivational excitement) and directional (by progressively and selectively updating values for executed actions) control over behavior, several inconsistencies remain unresolved. First, we found that manipulations of DA during the feedback epoch that strongly influenced future left/right choice behaviors (i.e updated action-values) did not change latencies on future trials (i.e unaffected state-value). In other words, outcome related DA transmission is necessary and sufficient to update action-values, but appears not to be involved in state-value update. This is a surprising finding because natural rewards are capable of reinforcing the internal estimates of both action and state-values. Does this mean that feedback-related DA-RPE deflections reflect some, but not all aspects of the reinforcing characteristics of natural rewards? Or, are state and action-values eligible for update by DA during different epochs within a trial?

Second, manipulations of DA at trial start that influence latency on the same trial, did not affect neither choice nor latencies on future trials. In other words, DA stimulation during this epoch has an immediate effect on behavior, but is not involved in updating values for future performance. This is consistent with other studies that find reinforcing

qualities of DA optogenetic stimulations restricted to epochs surrounding reward outcome (Steinberg et al., 2013; Chang et al., 2016). Does this suggest a temporal specificity for the causal action of DA in mediating action-reinforcement? Indeed, this notion is supported by in-vitro studies reporting that DA induces cellular and synaptic reorganizations (Shen et al., 2008; Surmeier et al., 2007; Kreitzer and Malenka, 2008) that are most potent during a narrow window of DA and glutamate coincidence (Yagishita et al., 2014).

In this chapter I assess the temporal organization of DA's control of behavioral learning. Specifically, I examined how rapid DA fluctuations within a trial contribute to the update of state and action-values to respectively affect future vigor and choice behaviors. To address this, rats performed the trial-and-error task with manipulations of feedback cues or optogenetic stimulations of the VTA. Based on our initial observations of DA stimulations during the feedback epoch, we hypothesized that the plasticity window for DA mediated state-value update is different from this outcome epoch (perhaps pellet-consumption epoch). We further hypothesized that DA mediated action-value reinforcement is restricted to the feedback epoch.

Contrary to our first hypothesis, we found a strong behavioral window of state-value update during the feedback epoch. Furthermore, variably timed optogenetic VTA stimulations (that influence choice on future trials) did not affect the update of state-values, suggesting that mesolimbic DA is not sufficient (on its own) to update state-values. By contrast, we found that VTA stimulations strongly favored choice re-selection, but, this effect of DA stimulation was restricted to the feedback epoch. Together our findings elaborate on the causal involvement of DA in mediating value-reinforcement, indicating a divergent role of DA-feedback signals for state and action-value modifications.

Methods

Trial-and-error task

The trial-and-error task (Figure 3.1a) used here is identical to that described in chapter 2, with some modifications. Briefly, each trial started with the illumination of only the central port, awaiting instrumental responding. Rats must enter and maintain the center-poke for a variable hold duration of 0.5-1.5 seconds, followed by a 250ms long white noise 'GO' cue. Rats subsequently made rapid free-choices to the rightward or leftward adjacent ports. If a trial is rewarded, a sugar pellet is dispensed at the same time rat pokes side-port, generating an audible reward cue ('click' from food hopper). In unrewarded trials, breaking the side port beam omitted reward without any audio-visual cue. Exiting out of the side port marked the end of a trial and a random ITI (6-10 seconds) was initiated. Entry into non-illuminated ports to start a trial (wrong starts) or violation of hold-duration (premature movement before 'GO' cue, classified as false starts) caused the house-light to turn on, aborting trial and initiating the ITI. To promote cued responding at the beginning of each trial, center-port illumination was delayed by 0.5 seconds if rat response was detected at any of the 5 ports. Left and right choices had independent reward probabilities, each maintained for blocks of 35-45 trials (randomly selected block length and sequence for each session). All combinations of 10%, 50% and 90% reward probability were used including 10:10 and 90:90. Training to perform the trial-and-error task typically took 6-10 weeks, and included several pretraining stages as described in chapter 2 (advancing when ~85% of trials were performed without procedural errors).

To assess if the reward-cue is necessary and/or sufficient to communicate signals critical for learning, we systematically decoupled reward-cues from the reward itself. To test if the reward cue is sufficient to update motivational values, we provided the reward cue on 30% of randomly selected trials that did not receive a sugar pellet reward ('fake' reward-cue trials). We analyzed latency distributions following these trials to assess if the reward-cue alone (i.e. without pellet consumption) can update internal estimates of

reward expectation. Next, to test if the reward-cue was necessary for learning/updating of values, we omitted the playback of the the reward-cue on 30% of randomly selected rewarded trials ('absent' reward-cue trials). Rats received the sugar pellet at the food port. Next-trial latency was analyzed in the same manner to assess if the reward-cue is required to promote strengthening of values. 8 total rats were exposed to the modified version of the trial-and-error task, each performing a single session of both types.

Surgery

To achieve cell-type specific expression of channelrhodopsin (ChR2) selectively in dopaminergic cells, we combined *TH-cre+* rats with CRE dependent viral constructs. Under isoflurane anesthesia, 12 rats were injected with AAV5-EF1α-DIO-ChR2-EYFP (University of North Carolina vector core, 6.3 x 10^{12} viral particles ml^{-1}, 1.5uL per injection site) bilaterally into the VTA (coordinates from bregma: -5.2mm AP, 1.0mm ML, 7.3mm DV from dura). Pulled glass micropipettes pressure-ejected virus into the midbrain at 50 nL/min and needle was left in place for 5 minutes to allow diffusion. After injection, 2.5 mm optic fiber ferrules (200um core diameter) were bilaterally cemented ~200um dorsal to injection site using dental acrylic and skull screws. We verified viral expression and fiber placement in post-mortem tissue processed for fluorescent microscopy.

Optical manipulation during behavior

Following surgeries, rats received postoperative care and were allowed to recover for one week. Optogenetic testing was performed exactly as described in chapter 2. For all behavioral testing (starting 5 week after virus injection), a Casio blue laser diode (445nm) was TTL modulated at 30Hz (15 pulses, each 10ms wide) for a total pulse train of 0.5 seconds at a fixed power (20mW at patch cable tip). This particular combination of parameters was previously verified to produce striatal DA levels comparable to those seen in response to unexpected reward delivery (Figure 2.6).

Behavioral/optogenetic testing was organized into two experiments. First, rats received only one of five stimulations triggered by the following events: (i) Light-On, (ii) Center-In, (iii) 'Go' cue, (iv) Side-In, and (v) Food-portIn. Five rats participated in types i,iii,iv and v. An additional seven rats (who performed experiment 2) received Center-In illumination experiments (Table 3.1). For these experiments, optical stimulation was triggered by a specific behavioral event on a randomly selected 30% of trials. Each rat performed three sessions, with an average of 380 +/- 23 trials per session. To assess the causal role of DA in reinforcing choice behaviors (and how it changes within trial), we quantified the probability of left/right choice re-selection on trial n+1, if stimulation was delivered on trial n. We have previously reported (chapter 2) that reward receipt strongly affects probability of repeating (p(repeat)) action on the next trial. Thus, we first separated trials into rewarded/unrewarded, and compared how p(repeat) changes for stimulated and unstimulated trials. We averaged results from each session, for each rat, and performed two-way repeated measures ANOVAs to assess the main effects of LASER and REWARD. All p-values reported are Bonferroni adjusted for multiple comparisons.

Experiment	# of sessions per rat	RAT ID						
Light-On STIMULATION	3	IM-529	IM-531	IM-532	IM-533	IM-534		
Center-IN STIMULATION	3	IM-617	IM-618	IM-619	IM-620	IM-641	IM-642	IM-643
GO' cue STIMULATION	3	IM-529	IM-531	IM-532	IM-533	IM-534		
Side-In STIMULATION	3	IM-529	IM-531	IM-532	IM-533	IM-534		
Foodport-In STIMULATION	3	IM-529	IM-531	IM-532	IM-533	IM-534		
Variable STIMULATION	5	IM-617	IM-618	IM-619	IM-620	IM-641	IM-642	IM-643

Table 3.1: Breakdown of rats used for various experiments.

Each row represents one experimental condition. Behavior aligned stimulations are summarized in the first 5 columns. Experiment 2 (described below) made use of 7 rats that each performed 5 sessions.

In the second experiment, we decoupled stimulation from behavior, and delivered optical stimulation randomly jittered from trial start. Each trial had a 30% likelihood of receiving stimulation (same 30Hz pulse train), and the timing of laser stimulation was randomly selected from a uniform distribution of 0-5 seconds (1ms precision) relative to Center-In. Seven rats performed the trial-and-error task (Table 3.1), each completing 5 sessions (371 trials +/- 9 SEM per session). To assess the influence of variably timed laser stimulations on behavior, we first combined data from all rats (12,977 total trials, 8,824 unstimulated and 4,153 stimulated trials), re-aligned stimulation events relative to all task events (Figure 3.3), and evaluated p(repeat) in a sliding 100ms analysis bin. Any 0.5 second long laser stimulation was defined to fall in the analysis bin if its edges cross the boundaries of the current bin. This way, each stimulation event contributes to a maximum of five consecutive analysis windows. Bins that had less than 50 stimulated trials were omitted from further analysis. We stepped the analysis bin across all alignments in 100ms steps (no overlap), and plotted the laser-induced additional reinforcement (i.e difference between mean p(repeat) in current bin and mean unstimulated p(repeat)). To test if observed laser induced additional reinforcement was significant, we used a shuffle test for each bin. We shuffled the trial type label for all control and stimulated trials of current bin 10,000 times, and for each shuffle we computed the laser-induced reinforcement. To obtain a p-value, we evaluated the fraction of shuffles with mean reinforcement larger (or smaller) than the actual reinforcement value of the current bin. To determine the critical p-value that is corrected for multiple comparisions, we used the Benjamini and Hochberg's correction for false discovery rate (Benjamini and Hochberg, 1995). Briefly, all p-values for an alignment with n-bins were sorted in ascending order, and the largest order index (i) for which $p(i) < (0.1*i)/n$ served as the critical threshold for hypothesis rejection.

Results

Latency is selectively sensitive to state-value update

We have previously (chapter 2) demonstrated that during trial-and-error task performance, rats readily adapt their motivation for task-engagement in response to changes in reward rate. To get a handle on the precise mechanism of DA-mediated learning of state-values, we first examined how natural rewards modulate trial-by-trial latencies.

Latencies to initiate trials were longer for low-reward blocks in contrast to shorter latencies for high-value blocks (Figure 3.1a), indicating modulation by the rate of reward (Wang et al., 2013). The observed variability in latency was more strongly correlated to overall block state-value than the nominal value for the subsequently chosen action (Figure 3.1b,c), indicating that latency is an index of overall reward rate, and not strongly influenced by action-specific outcomes. To directly compare whether variability in latency is better captured by changes in state or action values, we performed a multiple regression, with latency as independent variable, and both combined block-value and chosen-arm value as predictors. We found that block state-values accounted for a significantly larger variability in latency than block arm values (paired Mann-whitney test of regression coefficients, $p < 0.001$). By contrast, another behavioral metric, the reaction time (RT, time from 'GO' cue to Center out) was selectively responsive to changes in values for the chosen option (Figure 3.1d,e). Furthermore, a significantly larger fraction of RT variability was captured by changes in chosen arm value than net block-value (paired Mann-whitney test of regression coefficients, $p < 0.001$). Thus, latency and RT reflect the behavioral expression of separable valuation mechanisms for overall rate-of-return and action-specific values respectively.

Prior investigations assume inherently slow dynamics in the fluctuations of motivated behaviors, or brain (and peripheral) circuits that control motivational arousal (e.g. tonic DA levels or circulating hormones). On the contrary, we find that motivation to

perform the task can adapt to changes in rewards on a trial-by-trial basis. That is, the latency change in adjacent trials (i.e trial-by-trial variability) is strongly modulated by reward-surprise. Specifically, earning an additional reward in low reward-expectation trials (i.e very surprising reward) cause a speeding of latency on the next trial by several seconds. This effect is progressively smaller for less surprising rewards (Figure 3.1f, red circles). By contrast, omission of a highly expected reward causes a significant slowing of latency on the very next trial (Figure 3.1f, blue circles). Similarly, trial-by-trial changes in RT also adapt to reward surprise; RT is strongly reduced for the omission of expected reward, whereas earning an unexpected reward causes the speeding of RTs. These results indicate that state and action values can undergo single-trial updating, and that the behavioral expression of learned values display a recency-weighted modulation.

Feedback reward-cue is necessary and sufficient to update state-values

During performance of the task, a reward cue (an audible 'click') signals reward delivery upon side-port entry. This reward cue is the major feedback signal, and we have previously reported robust DA changes that scale with RPEs during this epoch (Figure 2.5a,b). Thus, If rats continuously modulate their motivation to work via dynamic updating of state values, are these reward-cues causally involved in state-value learning trial-by-trial?

To test if the the reward-cue was *sufficient* to promote state-value reinforcement, we exposed 8 rats to a modified version of the trial-and-error task. In this version, 30% of trials received an identical reward-cue, but we omitted the sugar pellet delivery into the food receptacle (hence, 'fake' reward-cue). Under these conditions, we can assess if state values can undergo reinforcement solely by reward feedback, or alternatively, if consuming the sugar pellet is required for state-value update.

To ensure that rats did not extinguish the meaning of reward-cues (i.e. association to reward), each rat performed only one session, with trials randomly selected to receive the 'fake' reward-cue. We found that rats did not distinguish between real or 'fake' reward-cues, as evidenced by equivalent haste to retrieve the sugar pellet at the food-

port (Figure 3.2a, $p = 0.13$, two tailed t-test). To examine if state-values underwent updating for the two reward-cue types, we assessed the latency distributions on the very next trial. Naturally rewarded trials (real reward-cue, in addition to sugar pellet delivery) were followed by latencies that were significantly shorter than unrewarded trials (no reward-cue and omitted sugar pellet, red and blue distributions, Figure 3.2b mean difference = 2.27 sec, $p < 0.001$).

Strikingly, trials that received a 'fake' reward-cue (without actual delivery of a sugar pellet) displayed fast latency distributions that were similar to rewarded trials (Figure 3.2b, orange distribution, mean difference = 0.17 sec, $p = 0.44$). Despite omission of the sugar pellet, observed latencies on these trials were significantly different from unrewarded trials (mean difference = 2.43 sec, $p < 0.001$), suggesting that consumption of a pellet is not involved in state-value update. In addition, 'fake' reward-cues speeded latencies with an similar recency-weight as real reward-cues (Figure 3.2c). Together, these results indicate that evaluative feedback of the reward-cue is sufficient to update state-values and their subsequent behavioral expression.

Next, we tested if the reward-cue was *necessary* for state-value update. To examine this, 8 rats performed another variant of the trial-and-error task where 30% of rewarded trials did not signal a reward-cue ('absent' reward-cue). Critically, on these 'absent' reward-cue trials, a sugar pellet was dispensed when the rats entered the food-port. In this manner, we can quantify latencies following trials where rats consumed a sugar pellet, but did not receive a reward-feedback. Latency analysis was restricted the subset of 'absent' reward-cue trials where rats retrieved the the sugar pellets (> 80%). Indeed, rats approached and retrieved the sugar pellets at the food-port with similar speed as normally rewarded trials (Figure 3.2d).

'Absent' reward-cue trials produced latency distributions that were significantly different from normally rewarded trials (Figure 3.2e, mean difference = 2.47 sec, $p < 0.001$), consistent with the notion that consumption of a pellet is not involved in state-value update. These 'absent' cue trials displayed latency distributions that were similar

to unrewarded trials (mean difference = 0.05 sec, 0.90), establishing the necessity of the reward-cue to update state-values.

Collectively these results support the notion that i) latency is a selective behavioral expression of learned state-value, ii) state-values are updated by recency wighted natural rewards, iii) latencies can adapt to surprising (or disappointing) rewards within a single trial, iv) and finally, evaluative feedback signaled by the reward-cue during side-port entry are both necessary and sufficient to update state-values.

Reward-related feedbacks recruit many brain systems, including the DA. Midbrain DA cells (and subsequent forebrain release) are extensively documented to become modulated during the feedback epochs to signal RPEs, used for incremental updates of reward-related values. Thus, rapid DA release is argued to signal the reinforcing quality of natural rewards, capable of strengthening both state and action values. Below, we tested whether artificial DA stimulations are sufficient to update state and action-values, and assed the time course underlying these DA mediated reinforcement.

VTA stimulations during different epochs do not affect future-trial RT and latency

If DA provides the reinforcement signal for state-value update, we hypothesized that activation of the VTA during the outcome epoch would be sufficient to affect latency on the next trial, even in the absence of sugar pellets. Alternatively, DA signaling during a different epoch may be involved in state-value reinforcement. A second alternative is that DA signaling may not be sufficient (by itself) to reinforce state-values, and may function as part of a larger learning circuit to mediate changes in value.

Our previous optogenetic manipulations suggested that DA manipulations during the feedback epoch is neither necessary nor sufficient (Figure 2.14) to affect next-trial latency. Replicating our previous findings in a new cohort of rats, optogenetic stimulation of the VTA during the side-in epoch did not affect latency on the next trial (Figure 3.3d), indicating that these stimulations did not strengthen stored state-values. To assess if DA during other epochs was critical for state-value update, we delivered

VTA stimulations coincident with various behavioral events during task performance. Providing stimulations at Light-on, Center-in, 'GO' cue, or Foodport-in did not affect latency on the next trial (Figure 3.3 a-e). This indicates that DA is not sufficient (by itself) to update state-values, and that task related striatal DA release events (Figure 2.10), are not causally involved in the update of state-values.

RT changes trial-by-trial are also sensitive to outcomes of the very last trial (Figure 3.1g). That is, earning a reward significantly speeds the latency toward the chosen action for the next trial, reflecting the vigorous execution of learned action value. To assess if laser stimulation affected RT on the next trial, we separate RT distributions for stimulated and unstimulated trials selectively for trials the animal repeated its choice. In this manner, we can assess if laser stimulations that produced an increased likelihood of repeating a choice also produce changes in RT. We found that next-trial RTs were not significantly reduced (Figure 3.3 f-j), suggesting that DA stimulation did not affect the vigorous expression of learned action-values.

DA updates action-values selectively during the feedback epoch

Next, we turned to the causal involvement of DA in mediating choice reinforcement (i.e action-value update). We quantified how the probability of action reelection (p(repeat)) is influenced by VTA stimulations. We have previously found that DA stimulation during the feedback epoch is causal to action-value update, bidirectionally modulating action re-selection (Figure 2.6). We extended these optical stimulations to different epochs within the trial to assess the temporal characteristic of this reinforcement window. We hypothesized that this window would be restricted to the feedback epoch.

Figure 3.4 summarizes changes in p(repeat) in experiment-1 for each stimulation condition, broken down by reward. Replicating our previous finding in a new cohort of rats, Light-On stimulation on trial n did not affect the likelihood of making the same choice on trial n+1 (Figure 3.4a, two-way ANOVA, LASER main effect $F(1,4) = 0.21$, $p = 0.67$, REWARD main effect $F(1,4) = 27.18$, $p= 0.006$). Furthermore, VTA stimulation

triggered by Center-In also did not produce choice reinforcement (Figure 3.4b, LASER main effect, $F(1,6) < 0.001$, $p = 0.91$, REWARD main effect $F(1,6) = 299.5$, $p < 0.001$). This half-second stimulation triggered by Center-in would elevate DA during the 'hold' period, where we reported a consistent volumetric plateau (Figure 2.3). Together, these results indicate that natural early-trial [DA] increases that correlated with approaching the center port are not sufficient to reinforce left/right actions that were executed shortly after.

Moving closer to epochs where rat executed the left/right choice, optical stimulation coincident with the 'GO' cue increased p(repeat) significantly (Figure 3.4c, LASER main effect $F(1,4) = 8.25$, $p = 0.04$, REWARD main effect $F(1,4) = 10.73$, $p = 0.031$). Further, as previously reported in chapter two, providing stimulation at Side-In significantly strengthened the action's association with reward, increasing the probability of choice re-selection (Figure 3.4d, LASER main effect $F(1,4) = 8.04$, $p = 0.04$, REWARD main effect $F(1,4) = 15.62$, $p = 0.016$). Lastly, Food port-In stimulation was associated with a slight elevation in repeating a choice, but was not statistically significant (Figure 3.1e, one-way ANOVA, LASER main effect $F(1,4) = 4.42$, $p = 0.103$).

Collectively, these results suggest that the causal involvement of DA in reinforcing choice behavior is temporally organized (Figure 3.5). That is, optical stimulations at Light-On or Center-In that are delivered earlier (relative to left/right action itself) do not strengthen its reward value used in future decisions. On the contrary, stimulation during the generation and execution of movement robustly enhanced action-value, and promote the re-selection choice. This plasticity window is apparent when we quantified the additional reinforcing quality of the optogenetic stimulation, or the difference between stimulated and unstimulated likelihoods (Figure 3.5). Laser effects did not show a strong interaction with rewards: none of the two-way ANOVAs for the various stimulation conditions yielded a significant LASER x REWARD interaction (data not shown).

While results from experiment 1 define a strong, but, broad DA mediated plasticity window, it is still difficult to ascertain the detailed temporal characteristics of this window

relative to task behavior. To gain an even closer insight into DA mediated reinforcement in relation to behavior, we performed a second experiment with variably-timed optical manipulation in the same session. Experiment two consisted of 35 sessions (from 7 rats, each performing 5 sessions) wherein 30% of trials were randomly selected to received the same 0.5 second pulse train, but stimulations were randomly delayed 0-5 seconds from Center-In (Figure 3.6b). To quantify the role of each variably timed stimulation in reinforcement, we re-aligned laser events relative to each behavioral event (Figure 3.6c), and computed the additional laser-induced reinforcement of choice behavior in 100ms moving windows.

In agreement with results from experiment 1, we primarily found strong windows for DA-mediated behavioral reinforcement surrounding the preparation and execution of choice behaviors (Figure 3.7). Specifically, we found a statistically significant laser induced reinforcement beginning at ‘Go’ cue both for rewarded and unrewarded trials. This additional reinforcement reached peak levels just after the Center-Out event, and for unrewarded trials, remained elevated until rats poked the side port.

Rewarded and unrewarded trials differed in the extent DA stimulation supplemented action-value learning. The overall time course of artificial DA reinforcement was greater for stimulation on unrewarded trials than rewarded trials. This likely reflects a a ceiling effect in rewarded trials, of how much additional strengthening synapses undergo when natural and artificial DA increases coincide. On the other hand, following omission, striatal DA at Side-In rapidly declines, and optical activation of the VTA may elevate DA levels to those comparable to rewarded trials.

Discussion

Forebrain DA levels can modulate the immediate and future expressions of learned values. DA's influence over ionic currents and cellular-excitability likely underlies immediate performance effects. On the other hand, DA can also affect future behaviors via induction of persistent synaptic plasticity in circuits for adaptive decision-making. In this chapter, I examined the precise role of DA in mediating behavioral learning of state and action-values.

I specifically demonstrated that during performance of the trial-and-error task, rats make use of two dissociable (but related) valuation mechanisms for specific left/right choices (action-values) and overall rate-of-return (state-values). Updated state and action-values behaviorally manifested as changes in trial-by-trial latency and reaction times respectively. The progressive update of these values, and their subsequent behavioral expression illustrate mechanisms for how rewards refine actions for goal-directed behaviors, in addition to exerting overall arousal/activational effects.

Motivational drives such as those behind seeking food, water or other behaviors sometimes change slowly, and thus, brain motivational mechanism are also suggested to inherently vary slowly. For example 'tonic' DA has been hypothesized to regulate motivational arousal, and despite the sluggish rate of change suggested by its name, the actual time-constant of extra-synaptic DA fluctuations are yet to be clearly documented. Nonetheless, we provide behavioral evidence for fast motivational changes in response to reward surprise. Specifically, we found that trial-by-trial latency and RT undergo single-trial (bidirectional) adaptations in response to reward feedback cues, suggesting that rapid motivational learning can be induced via fast signals for evaluative feedback.

I further examined how DA fluctuations influence learning of state and action-specific values. Initial optogenetic manipulations described in chapter two suggested that 1) DA is necessary and sufficient to update action-values to affect future-trial choices, 2) This effect of DA on action-value update may be restricted to the feedback

epoch, and 3) outcome related DA fluctuations do not update state-values. Elaborating on these findings, I report here that mesolimbic DA changes throughout the trial are not sufficient to update state-values as reflected in latency changes. Furthermore, the same optogenetic stimulations (that did not influence future latencies) reinforced choice behaviors on the very next trial. Furthermore, DA's capacity to mediate choice reinforcement was restricted to a window surrounding the reward feedback. This is consistent with prior reports that operant responding is reinforced by rewards (or brain-stimulation reward) within a narrow time window (Thorndike, 1911; Black et al., 1985; Dan and Poo, 2006). Additionally, studies that optically stimulate genetically identified DA cells have reported that the causal involvement of DA in action-reinforcement and stimulus-associations are restricted to epochs surrounding reward feedback (Steinberg et al., 2013; Chang et al., 2016).

The plasticity window for DA mediated action-values are consistent with prior reports of electrophysiological activity patterns of striatal MSNs. Specifically, dorsal striatal MSNs are active during the preparation and execution of left/right movements in similar operant tasks (Gage et al., 2010; Ito and Doya, 2015). In addition, these cells found in the medial-lateral axis of the dorsal striatum are reactivated during the feedback epoch. Together these observations suggest that DA release coincident with the activity of postsynaptic cells that code for action-values underlie functional plasticity to promote the behavioral re-selection of actions. While we initially predicted that this window would be narrow, and restricted to the feedback epoch, the observed extent of its temporal spread was unexpected, starting at the 'GO' cue and extending until food-port visit. This potentially suggests redundant brain mechanisms for action-value reinforcement. For example some parts of this window may be contributed by the effects of DA in the DMS (cells that are active during decision/action-planning) whereas other portions may be mediated by effects in DLS (cells that are active during movement execution). This, however, is speculative, and direct striatal manipulations of DA are required to test these predictions.

Rather paradoxically, DA stimulations that increased the selection likelihood of specific actions were not sufficient to cause changes in RT for those choices. This suggested that changes in synaptic strengths for updated action values are not reflected in the behavioral vigor of executing those actions. Furthermore, in contrast to clear action-value modulation by DA, state-values appeared insensitive to DA manipulations. We have behaviorally demonstrated that state-values (indexed by latency changes)are capable of undergoing immediate update by reward-feedback. Together these findings indicate that neural circuits for learning/updating values, arbitrating/selecting among valued options, and their (vigorous) execution may subtly differ in their sensitivity to DA transmission. For example, DA may be sufficient to induce changes in plasticity of circuits that are important for choice arbitration, as evidenced by an immediate strengthening of choice behavior on the next trial. However, the downstream circuit that executes the choice may require larger DA increases, or other conditions to produce the vigorous of execution of the selected options. Nonetheless, some of our negative results have several alternative interpretations.

First, our DA stimulations may not be strong enough to engage plasticity in target regions to affect latency on the next trial. This possibility is less likely because we have shown that reward-cues that can modulate next-trial latency, naturally cause 40-60 nM phasic DA increase (Figure 2.3c). Furthermore, our optogenetic stimulations are titrated to induce ~50 nM DA increase in the NAcc (Figure 2.6b). Nonetheless, stronger optogenetic DA stimulation experiments are necessary to test this possibility.

Second, DA release in forebrain loci critical for state-value update may not be sufficiently large to modulate latencies on next trials. We targeted the mesolimbic DA system with optical fibers targeting the VTA cells. There is extensive documentation that DA within the ventral striatum is critical for motivated behaviors (Ikemoto and Panksepp, 1999; Nicola, 2007), and pharmacological suppression (or activation) of NAcc neurons or DA signaling within this region robustly influence approach behaviors (Ishiwari et al., 2004; Nicola, 2010). Thus, the ventral striatum appears to be the node for integrating overall motivation and effort allocation (Ikemoto and Panksepp, 1999; Salamone and

Correa, 2012). Indeed, a direct optogenetic stimulation of NAcc DA is necessary to assess this possibility, with the working hypothesis that DA within this region is sufficient to modulate immediate and future emission of motivated behaviors.

Third, in addition to robustly activating the DA axis, reward-feedback also recruits other brain systems critical for learning. These include the OFC (Critchley and Rolls, 1996; Tremblay et al., 1998), striatal MSNs (Hikosaka et al., 1989; Schultz et al., 1992), striatal cholinergic interneurons (TANs) (Aosaki et al., 1994; Apicella et al., 1991), and brainstem centers for fast sensory responses (e.g. PPN (Pan and Hyland, 2005; Norton et al., 2011)). In this chapter, I principally focused on the role DA plays in mediating reward-reinforcement, but these other circuits likely make critical contributions to learning from evaluative feedback. For example, TANs are reported to pause during the feedback epoch (Morris et al., 2004; Ravel et al., 2001), and several in-vitro studies have described cholinergic modulation of DA neurotransmission (Threlfell et al., 2012; Cachope et al., 2012) that result in the enhancement of phasic RPE signaling (Threlfell and Cragg, 2011; Cragg, 2006). TANs are, thus, argued to enhance learning during feedback epochs by setting the appropriate network dynamics for optimal learning of behaviorally relevant values (Franklin and Frank, 2015). Together these clues indicate that state-value updates may require activation of multiple (or all) neural systems that are activated in response to reward cues, and that striatal DA increases my not be sufficient on its own to reinforce future motivated behaviors.

Finally, the incentive-motivation theory (Berridge, 2007) of DA provides yet an additional alternative interpretation of our optogenetic results on reinforcement. This view emphasizes a selective role of forebrain DA levels in influencing immediate behaviors. Indeed, fast ‘phasic’ and slow ‘tonic’ levels covary with reward rate and predict subsequent vigor of task-engagement. We have also demonstrated that optogenetic stimulations can immediately affect the willingness to engage in task, affecting online investment of effort. Critically however, this theory asserts that DA is not involved in learning, and that behavioral results involving DA manipulations that are interpreted as learning-effects are misclassified. Specifically, if DA cells are stimulated,

the immediate incentive value of the cues (or rewards) are argued to become amplified, thus producing energized reward-seeking behaviors. Future vigorous reward-seeking behaviors, in turn, require additional boosts of DA coincident with, or prior to, their emission. Thus, our finding that DA stimulations do not affect future motivational performance is consistent with the incentive salience hypothesis. However, our results that show that the same stimulations update action-values are in sharp contradiction to this view, and others that emphasize performance-only effects of mesolimbic DA.

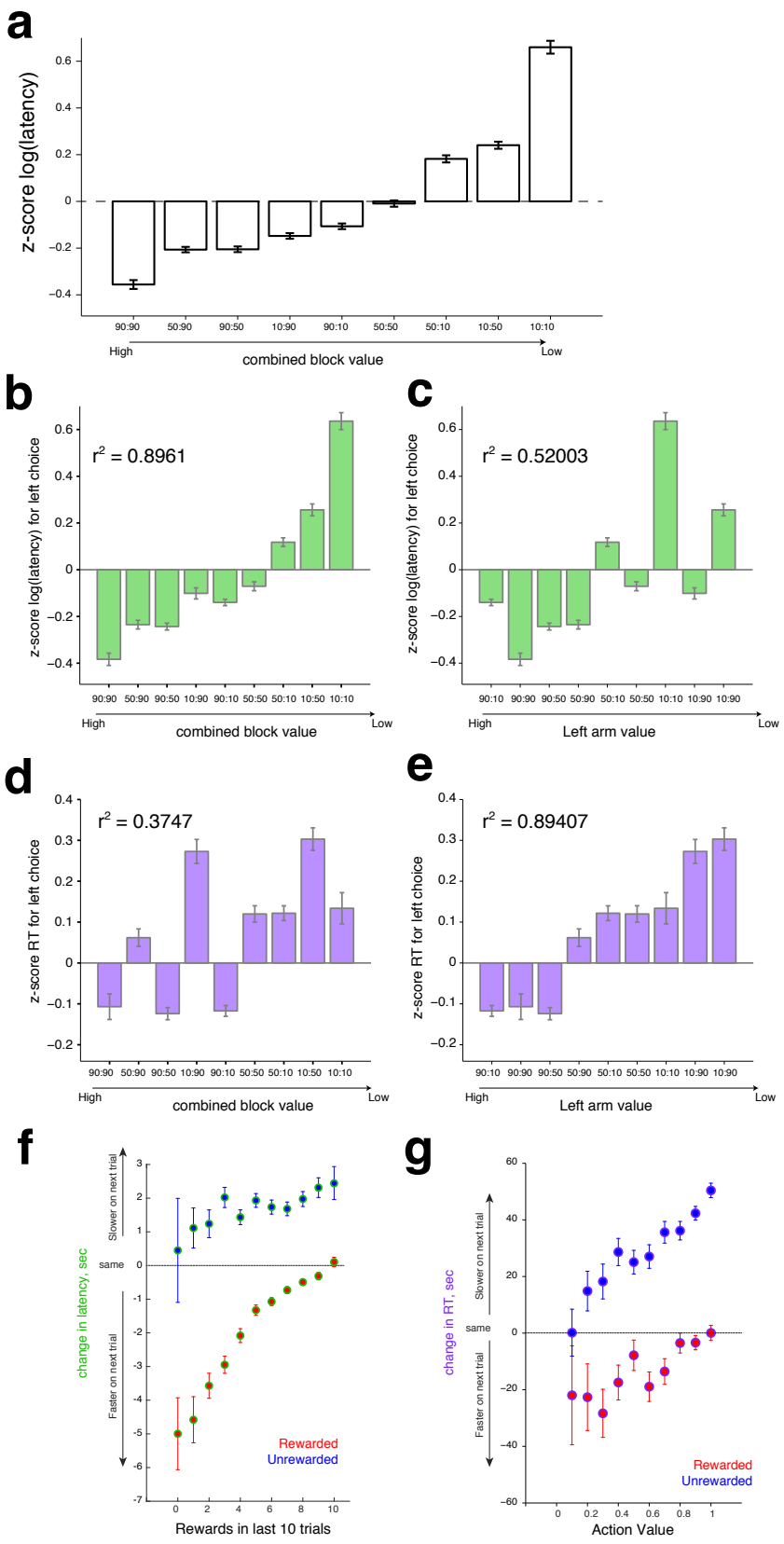

Figure 3.1: Latency is a selective behavioral expression of learned state-value

(a) Changes in normalized, log(latencies) a function of combined black value. Data from 16 rats, 13 +/- 2 sessions per rat. The same data is reanalyzed to assess how latencies in the second-half of blocks selectively on trials the rat would chose the left hand side are affected by combine block value (b) or the value of the left (chosen) arm (c). R-squared values are derived from linear regressions of latency and value. Same dataset and format for (d) and (e), but for reaction-times. (f) Changes in next trial latency depending on reward-surprise. (g) Adaptation in the RT to reward surprise. Here, value for the chosen action is estimated by assessing the fraction rewards among the last 10 choices for a particular side. We next assessed how the RT for that specific choice changed on the next trial depending on the outcome.

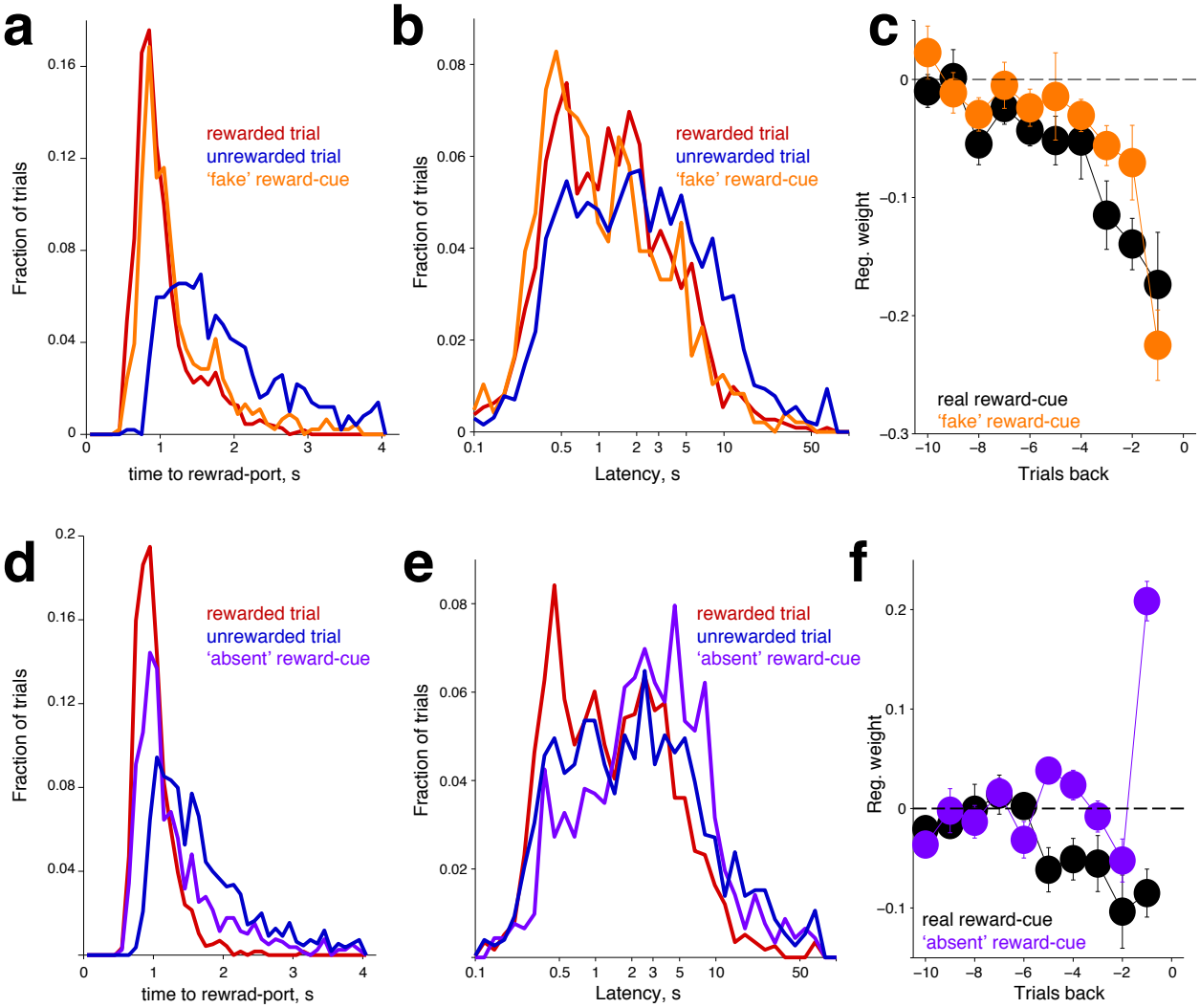

Figure 3.2: Feedback signals are necessary and sufficient to update state-values

(a) Distribution of time to food-port entry for rewarded, unrewarded and 'fake' reward cue trials. Data from 8 rats, each performing one session (371 +/- 12 trials per rat). Mean time to reward = 1.04 sec, 4.83 sec, and 1.10 sec respectively for rewarded (1120 trials), unrewarded (1289 trials), and 'fake' reward cue trials (491 trials). (b) Next-trial latency for the tree trial types. Mean latencies = 2.3 sec, 4.6 sec, 2.17 sec respectively for rewarded, unrewarded and 'fake' reward cue trials. (c) Regression coefficients of past, real or fake reward cues on latencies. (d) Data in the same format as (a), time to food port entry for rewarded (1170 trials), unrewarded (1531 trials), and 'fake' reward

cue trials (921 trials) . (e) Next-trial latency for the tree trial types. Mean latencies = 2.37sec, 5.23 sec, 5.18 sec respectively for rewarded, unrewarded and ‘absent’ reward cue trials.

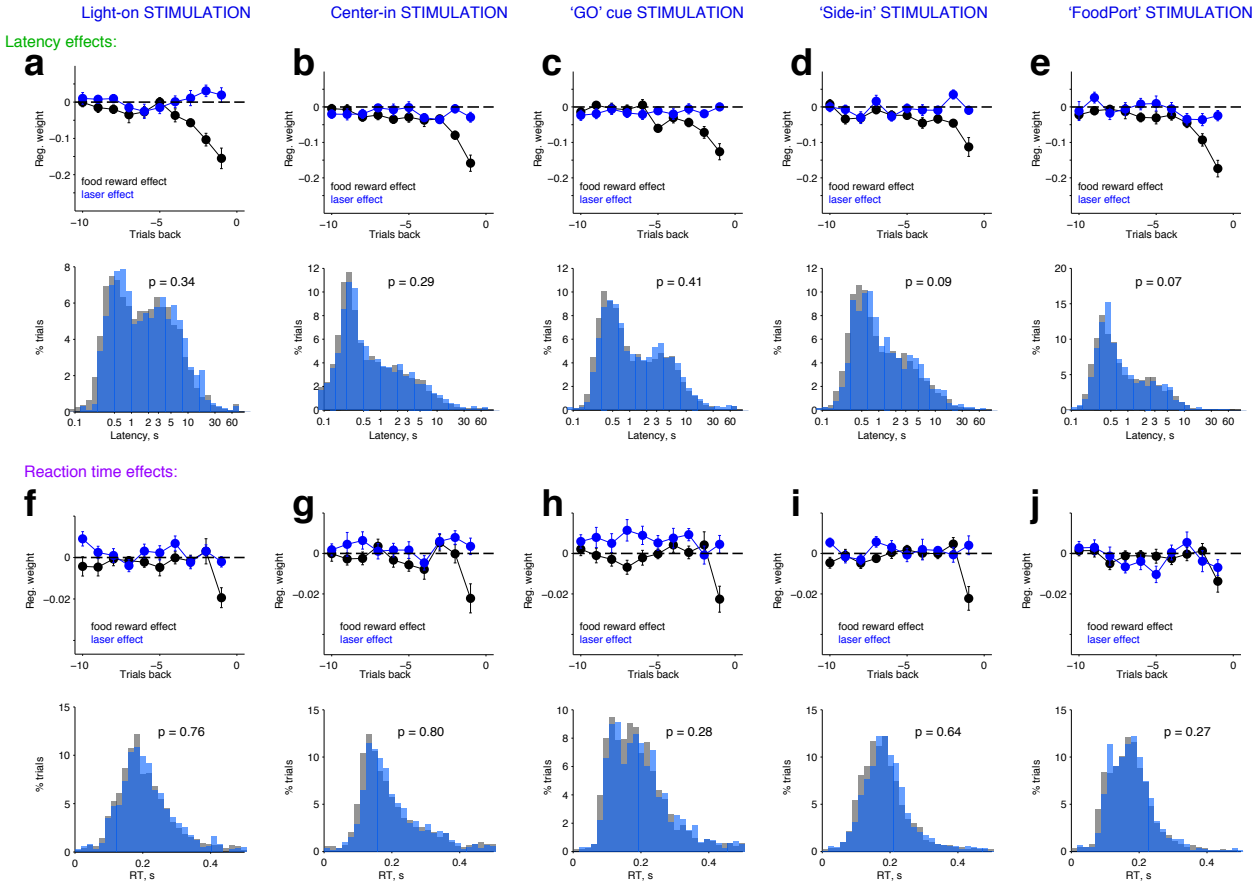

Figure 3.3: Variably timed VTA-stimulations do not affect future trial RT or latency

Quantification of variably timed optogenetic stimulations on latency effects (a) - (e), or RT (f) - (j), broken down by regression coefficients for the influence of rewards and stimulation (top), or distribution of the behavioral parameters on the next trial (bottom). All p-values are from two distribution Kolmogorov-Smirnov tests.

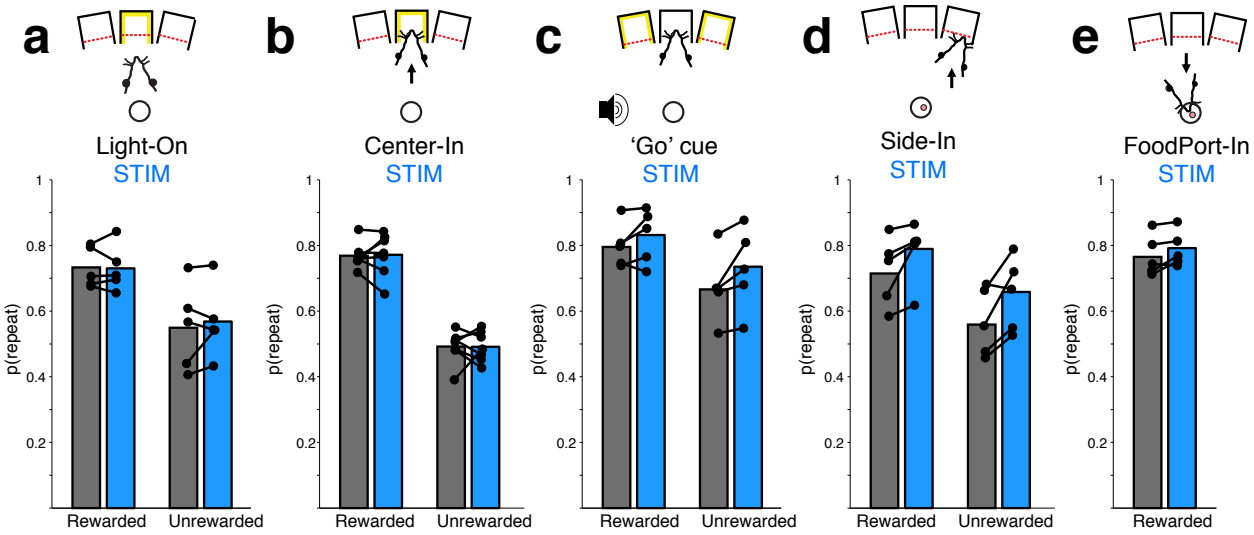

Figure 3.4: VTA stimulation surrounding outcome promotes choice reinforcement

Summary of changes in probability of repeating a choice. (a) Light-On stimulation did not affect choice re-selection on next trial. Data from 5 rats each performing 3 sessions, 363 trials +/- 12 SEM per session (b) Center-In stimulation also does not affect reinforcement. Data from 7 rats, 3 sessions per rat, 390 trials +/- 5 SEM per session (c) 'Go' cue stimulation significantly elevated the likelihood of choice re-selection. Data from 5 rats each performing 3 sessions, 379 trials +/- 7 SEM per session, (d) Side-In stimulation reinforces the chosen left or right action. Data from 5 rats, each performing 3 sessions, 385 trials +/- 13 SEM per session. (e) Food port In stimulation. Data from 5 rats, each performing 3 sessions, 403 trials +/- 5 SEM per session.

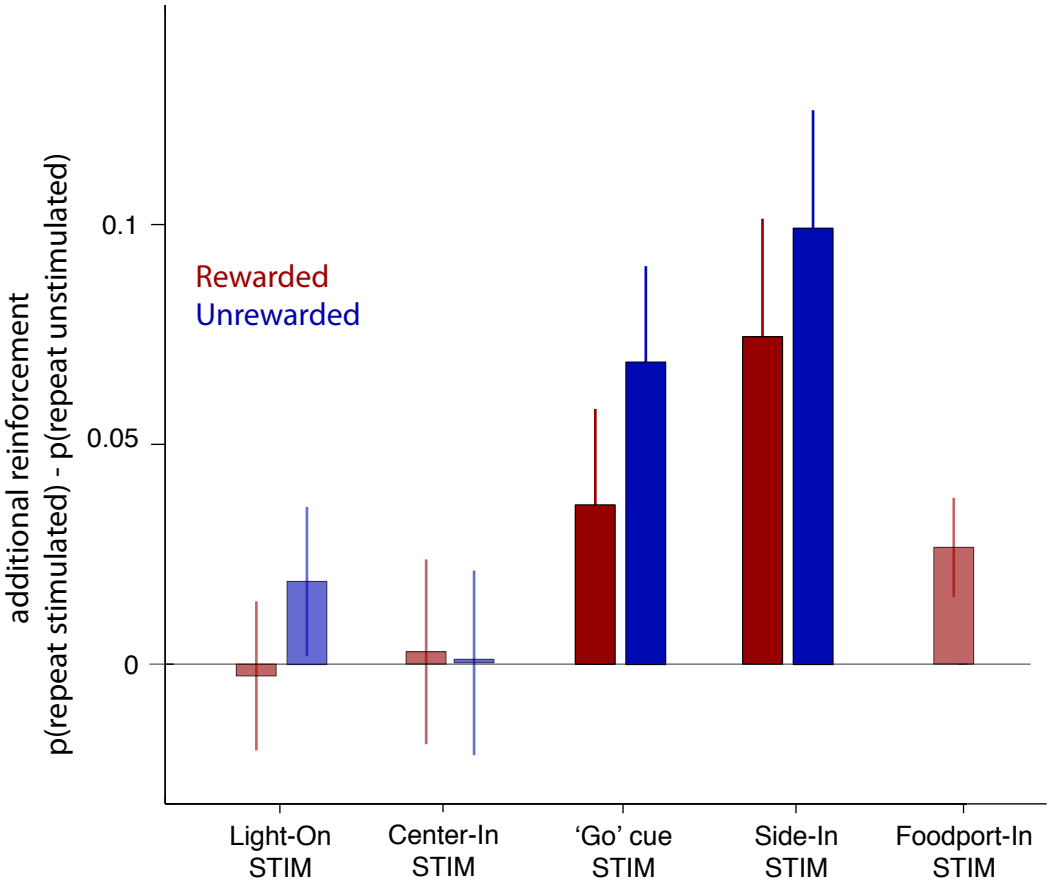

Figure 3.5: Summary of laser induced reinforcement

The difference between laser and control (respectively blue and gray bars in fig 3.2) were evaluated for rewarded and unrewarded trials. Statistically significant laser effects are shown in dark colors, while non-significant results are faded.

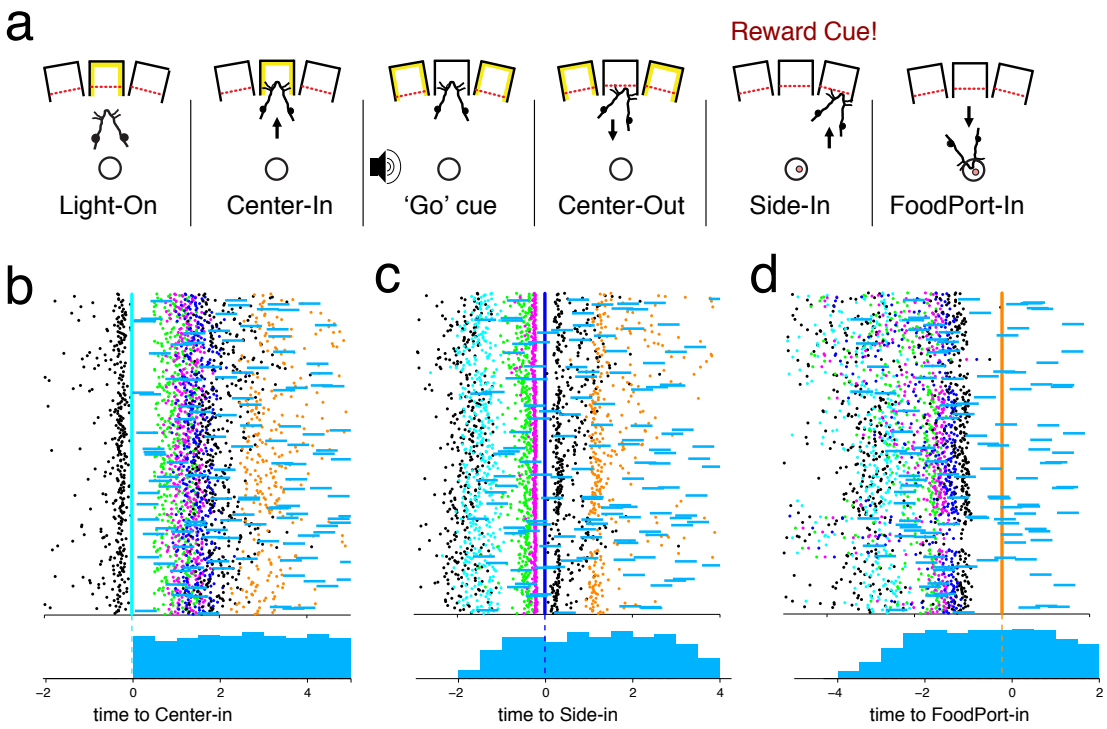

Figure 3.6: Experimental scheme for of variably-timed optical stimulations

The trial-and-error task is composed of a sequence of behavioral events (a) that appear with variable delay from each other during task performance. (b) Top, a representative example from one session from one rat displaying the temporal variability of behavioral events, and organization of laser stimulations. Events are shown aligned to Center-In (cyan color). Black dots to the left are Light-On events, green dots are 'Go' cue events, pink dots represent Center-Out events, blue dots are Side-in events, back dots are Side-Out events and finally orange dots are Foodport-In events. Blue dashes that appear in a subset of trials are 0.5 second long optical stimulation events, also aligned to Center in. Bottom, laser stimulation has a uniform distribution when aligned to Center-In. The same data displayed in (b) is re-aligned relative to (c) Side-In or (d) Food-port entry. The bottom panels show the resulting distribution of trials that received stimulation in the 0.5s bins relative to alignments.

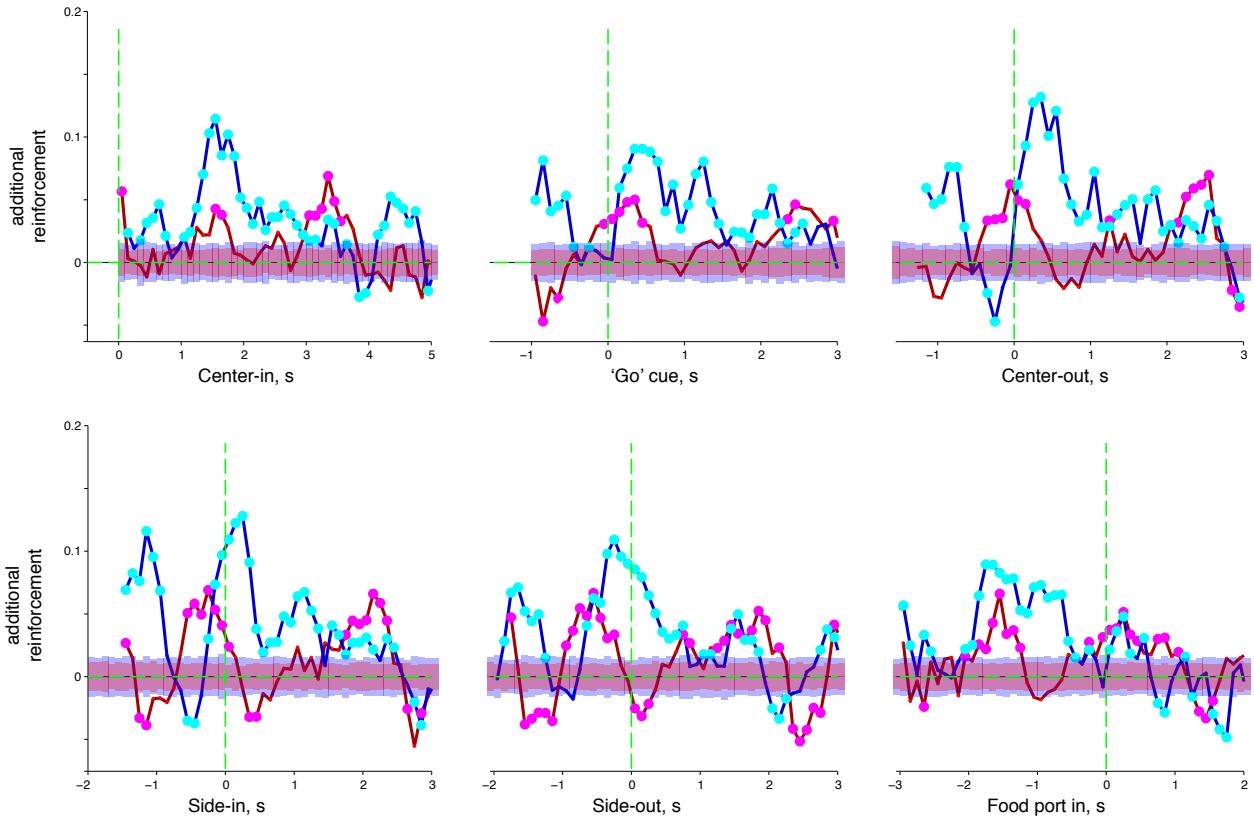

Figure 3.7: Reinforcement time course of variably timed VTA stimulations.

Summary of the time course of laser-induced reinforcement for variably-timed optogenetic stimulation. Blue traces quantify the laser effect for unrewarded trials, and red traces for rewarded trials. Each alignment plots results from 100ms sliding window, quantifying the effect of optogenetic stimulation delivered in the current bin relative to alignment. Cyan and pink dots mark data points with statistically significant observations (shuffle test with multiple comparison correction, see methods). Data is from 7 rats each performing 5 sessions, 371 trials +/- 9 SEM per session. In sum, 12,977 total trials, 8,824 unstimulated and 4,153 stimulated trials analyzed were analyzed. Blue and red transparent bars indicate 95% confidence interval for shuffle test, respectively for unrewarded and rewarded trials.

References

Adamantidis, A.R., Tsai, H.C., Boutrel, B., Zhang, F., Stuber, G.D., Budygin, E.A., Touriño, C., Bonci, A., Deisseroth, K., and de Lecea, L. (2011). Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. J Neurosci *31*, 10829-10835.

Aosaki, T., Tsubokawa, H., Ishida, A., Watanabe, K., Graybiel, A.M., and Kimura, M. (1994). Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. The Journal of neuroscience *14*, 3969-3984.

Apicella, P., Scarnati, E., and Schultz, W. (1991). Tonically discharging neurons of monkey striatum respond to preparatory and rewarding stimuli. Experimental brain research *84*, 672-675.

Berridge, K.C. (2007). The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) *191*, 391-431.

Black, J., Belluzzi, J.D., and Stein, L. (1985). Reinforcement delay of one second severely impairs acquisition of brain self-stimulation. Brain Res *359*, 113-119.

Cachope, R., Mateo, Y., Mathur, B.N., Irving, J., Wang, H.-L., Morales, M., Lovinger, D.M., and Cheer, J.F. (2012). Selective activation of cholinergic interneurons enhances accumbal phasic dopamine release: setting the tone for reward processing. Cell Rep *2*, 33-41.

Chang, C.Y., Esber, G.R., Marrero-Garcia, Y., Yau, H.J., Bonci, A., and Schoenbaum, G. (2016). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci *19*, 111-116.

Cragg, S.J. (2006). Meaningful silences: how dopamine listens to the ACh pause. Trends Neurosci *29*, 125-131.

Critchley, H.D., and Rolls, E.T. (1996). Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. J Neurophysiol *75*, 1673-1686.

Dan, Y., and Poo, M.-M. (2006). Spike timing-dependent plasticity: from synapse to perception. Physiol Rev *86*, 1033-1048.

Franklin, N.T., and Frank, M.J. (2015). A cholinergic feedback circuitto regulate striatal population uncertainty and optimize reinforcement learning. eLife , e12029.

Gage, G.J., Stoetzner, C.R., Wiltschko, A.B., and Berke, J.D. (2010). Selective activation of striatal fast-spiking interneurons during choice execution. Neuron *67*, 466-479.

Hikosaka, O., Sakamoto, M., and Usui, S. (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. J Neurophysiol *61*, 814-832.

Ikemoto, S., and Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Research Reviews *31*, 6-41.

Ishiwari, K., Weber, S.M., Mingote, S., Correa, M., and Salamone, J.D. (2004). Accumbens dopamine and the regulation of effort in food-seeking behavior: modulation of work output by different ratio or force requirements. Behav Brain Res *151*, 83-91.

Ito, M., and Doya, K. (2015). Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. J Neurosci *35*, 3499-3514.

Kim, K.M., Baratta, M.V., Yang, A., Lee, D., Boyden, E.S., and Fiorillo, C.D. (2012). Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PLoS One *7*, e33612.

Kreitzer, A.C., and Malenka, R.C. (2008). Striatal plasticity and basal ganglia circuit function. Neuron *60*, 543-554.

Maia, T.V., and Frank, M.J. (2011). From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci *14*, 154-162.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., and Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron *43*, 133-143.

Nicola, S.M. (2007). The nucleus accumbens as part of a basal ganglia action selection circuit. Psychopharmacology (Berl) *191*, 521-550.

Nicola, S.M. (2010). The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. J Neurosci *30*, 16585-16600.

Norton, A.B., Jo, Y.S., Clark, E.W., Taylor, C.A., and Mizumori, S.J. (2011). Independent neural coding of reward and movement by pedunculopontine tegmental nucleus neurons in freely navigating rats. European Journal of Neuroscience *33*, 1885-1896.

Pan, W.X., and Hyland, B.I. (2005). Pedunculopontine tegmental nucleus controls conditioned responses of midbrain dopamine neurons in behaving rats. J Neurosci *25*, 4725-4732.

Ravel, S., Sardo, P., Legallet, E., and Apicella, P. (2001). Reward unpredictability inside and outside of a task context as a determinant of the responses of tonically active neurons in the monkey striatum. The Journal of neuroscience *21*, 5730-5739.

Salamone, J.D., and Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. Neuron *76*, 470-485.

Schultz, W. (2007). Multiple dopamine functions at different time courses. Annu Rev Neurosci *30*, 259-288.

Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. The Journal of Neuroscience *12*, 4595-4610.

Shen, W., Flajolet, M., Greengard, P., and Surmeier, D.J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. Science *321*, 848-851.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci *16*, 966-973.

Surmeier, D.J., Ding, J., Day, M., Wang, Z., and Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends Neurosci *30*, 228-235.

Thorndike, E.L. (1911). Animal intelligence: Experimental studies (Transaction Publishers).

Threlfell, S., and Cragg, S.J. (2011). Dopamine Signaling in Dorsal Versus Ventral Striatum: The Dynamic Role of Cholinergic Interneurons. Front Syst Neurosci *5*

Threlfell, S., Lalic, T., Platt, N.J., Jennings, K.A., Deisseroth, K., and Cragg, S.J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron *75*, 58-64.

Tremblay, L., Hollerman, J.R., and Schultz, W. (1998). Modifications of reward expectation-related neuronal activity during learning in primate striatum. J Neurophysiol *80*, 964-977.

Tsai, H.C., Zhang, F., Adamantidis, A., Stuber, G.D., Bonci, A., de Lecea, L., and Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science *324*, 1080-1084.

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

Witten, I.B., Steinberg, E.E., Lee, S.Y., Davidson, T.J., Zalocusky, K.A., Brodsky, M., Yizhar, O., Cho, S.L., Gong, S., Ramakrishnan, C., Stuber, G.D., Tye, K.M., Janak, P.H., and Deisseroth, K. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron *72*, 721-733.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science *345*, 1616-1620.

CHAPTER 4: DISSOCIABLE SIGNALS FOR VALUE AND PREDICTION ERROR ON DOPAMINE MIDBRAIN-FOREBRAIN AXIS

Introduction

Electrophysiological studies of action potential discharge in midbrain DA cells are extensively documented to signal reward prediction errors (RPE) (Schultz, 1997; Schultz, 2016), both under pavlovian conditions (Cohen et al., 2012a; Fiorillo et al., 2008; Eshel et al., 2015; Eshel et al., 2016), and instrumental task performance (Bayer and Glimcher, 2005; Satoh et al., 2003; Roesch et al., 2007). On the other hand, measurement of forebrain DA concentrations (specially mesolimbic DA) using voltammetry and microdialysis indicate that DA conveys a motivational signal (Phillips et al., 2003; Roitman et al., 2004; Onge et al., 2012) - the temporally-discounted, estimated value of work (Hamid et al., 2016). The experimental and theoretical details of how fast fluctuations in DA code for this value signal is described in chapter 2. In addition, we provided evidence for how these DA value signals are simultaneously used for making decisions about whether to work (motivational decisions), and reinforcing executed actions in chapter 3.

A direct assessment of both DA cell spiking and forebrain release is, however, absent. Many DA studies make use of different tasks (or markedly different variants of the same task) or animal species (mice, rats or monkeys). Therefore, it remains elusive whether the quantitative decision-signal coded by VTA firing is the same decision variable that covaries with DA concentrations. In other words, is midbrain spiking disjoint from release of DA in striatum, and is such disparity in activity a reflection of an adaptive computational strategy?

While some have argued that activity patterns of midbrain DA cells are redundant, broadcasting a uniform message to target sites (Schultz, 1998; Glimcher, 2011; Eshel et al., 2016), many empirical reports provide evidence for the contrary. For example, DA cells form diverse groups based on biophysical characteristics (Lammel et al., 2008), anatomical projection-targets or input specificity (Ikemoto, 2007; Watabe-Uchida et al., 2012; Lammel et al., 2008; Lerner et al., 2015), and differentially recruited in response to salient reward or aversive cues (Cohen et al., 2012b; Matsumoto and Hikosaka, 2009). Additionally, forebrain DA release is under strong control of the local microcircuitry, via several receptors expressed on DA axons (Sulzer et al., 2016; Threlfell and Cragg, 2011). For example, tonically active cholinergic interneurons of the striatum (TANs) appear to regulate both slow and fast efflux of DA into the cleft via nicotinic and muscarinic receptor modulation (Rice and Cragg, 2004; Threlfell et al., 2012; Cragg, 2006). Furthermore, a striking recent study provided evidence for functionally-silent dopaminergic synapses whose vesicles do not undergo exocytosis under stereotypical release conditions (Pereira et al., 2016). Together, these observations establish mechanisms underlying potentially disparate signals of mesolimbic DA on the midbrain-forebrain axis.

Advanced simultaneous methods for assessing midbrain activity and striatal DA concentration are yet to be developed. Furthermore, combining existing techniques for assaying DA release and electrophysiology of VTA spiking is currently finicky and too difficult, although several groups are making efforts to develop integrated devices.

Quantification of changes in intracellular calcium levels has been used for many decades as a reliable method to detect action potentials in excitable cells (such as neurons and muscle cells) (Tsien, 1983), and exocytosis in glandular cells (e.g. pancreatic acinar and adrenal chromaffin cells) (Petersen and Ueda, 1976; García et al., 2006). In addition to being exquisitely regulated (especially in it's temporal and spatial spread), calcium directly couples electrical excitation to vesicular fusion, causing it to be a robust assay of both the electrical and neurotransmitter release activity.

Genetically encoded calcium sensors (variants of calcium chelating proteins) have undergone many iterations of directed evolution to make them ultra-sensitive and exhibit sub-second kinetics (Tian et al., 2009; Akerboom et al., 2012; Chen et al., 2013). The gathering and detection of fluorescence is also streamlined, either via ultra-precise single or multi-photon microscopes (Denk et al., 1990; Svoboda et al., 1997), ultra-light weight endoscopes (Helmchen et al., 2001; Flusberg et al., 2008) or tethered bulk fluorescence measurements using photometry (Cui et al., 2013; Gunaydin et al., 2014; Lerner et al., 2015; Parker et al., 2016).

In this study, we used fiber photometry (see methods for details on instrumentation and data acquisition) to assess DA dynamics on the midbrain-forebrain axis. We initially intended to inject CRE dependent viral construct to drive GCAMP6f expression in VTA DA cells, allow trafficking into striatal axon arbors, and simultaneously measure calcium dynamics in the cell-body (to assay spiking) and terminal fields within the striatum (as an indicator of presynaptic DA efflux). Unfortunately, we were unable to mitigate several technical challenges surrounding simultaneous measurement. Specifically, I was not able to consistently collect simultaneous measurements, and stratal photometry signals were sometimes qualitatively different among the rats. This variably is likely due to specific placement of optic fibers, while some rats did not exhibit robust striatal photometry signals. So, in this chapter, I will focus on results from the VTA of 7 rats performing 3 different tasks that provide clear comparisons to published reports.

First, we made recordings in rats during performance of the trial-and-error task to complement our previous voltammetry observations (Hamid et al., 2016). Second, we used a probabilistically-rewarded linear track running task which is readily comparable to voltammetric recordings on linear track (Howe et al., 2013). Finally, we trained rats on pavlovian conditioning task similar to those extensively used to study error coding of DA cells (Day et al., 2007; Oyama et al., 2010; Schultz et al., 1993; Tobler et al., 2005) and allocation of motivational value to conditioned stimuli (Flagel et al., 2011; Saunders and Robinson, 2012).

Methods

All animal procedures were approved by the University of Michigan Committee on Use and Care of Animals. Male *TH-Cre*$^{+}$ rats with a Long-Evans background (300-500g) were maintained on a reverse 12:12 light:dark cycle and tested during the dark phase. Rats were mildly food deprived, receiving 15g of standard laboratory rat chow daily in addition to food rewards earned during task performance. Training and testing for pavlovian task and trial-and-error task was performed in computer-controlled Med Associates operant chambers (25cm x 30cm at widest point) each with a 5-hole nose-poke wall, as previously described (Gage et al., 2010; Leventhal et al., 2012; Schmidt et al., 2013; Leventhal et al., 2014).

Pavlovian task

Rats were trained in the same operant chamber as the trial-and-error task (Chapters 2,3) under pavlovian conditions (Figure 4.2a). Each trial consisted of a conditioned stimulus (CS), followed by delivery of sugar pellet (US). To help rats distinguish the instrumental task and pavlovian-CS task performed within the same operant box, the house-light was illuminated for the duration of pavlovian task performance. Each trial started with the onset of a randomly chosen CS tone-bips (pulses of sound ON for 50ms, and OFF for 100ms) at 2 kHz, 5 kHz or 9kHz for a total of 2.5 seconds (Figure 4.2a, right). Each of the three tone frequencies were associated with reward probability of 0, 0.5 or 1.0. The relationship between tone frequency and reward probability was maintained constant for each rat, but varied across rats to decouple frequency dependent neural or behavioral responses. A fourth trial type included the unpredicted delivery of reward. Each session consisted of 160 trials with approximately 40 of each type randomly selected. If a particular trial was rewarded, a 45mg sugar pellet was dispensed after a 0.5 second delay following the offset of the CS tone, and initiating a random ITI(15 to 30 seconds). Overhead video acquired images at 30 frames per second, and the entry into the food receptacle was monitored by an infrared beam.

Animals were typically trained for 2-4 weeks on this task before testing. For each trial, latency to break food-port beam relative to CS onset was recorded to disk by a custom-written LABVIEW software. To further assess the dynamically varying probability and timing of conditioned responding, we quantified the continuous rate of anticipatory beam-breaks at food port (similar to anticipatory licking behaviors of monkey and mice in previous investigations (Cohen et al., 2012a; Eshel et al., 2015; Eshel et al., 2016). We tested if rats behaviorally (via anticipatory responses) classify CSs of different value using the the area under receiver operator curve (auROC) analysis. For each trial, we first counted the cumulative sum of time spent breaking food-port beam during CS playback, producing a distribution of each trial type. The trial-by-trial cumulative response during the 2.5 seconds preceding unpredicted reward was used as 'baseline' distribution. Across the three probabilistic trial types, the distribution of response durations was compared to the baseline distribution using standard ROC analysis for every rat. Statistical significance tests were performed using the shuffle test, where we first randomize group labels for each trial 10,000 times and evaluated auROC values. P-values were then computed as the fraction of shuffles larger (or smaller) than observed ROC values for non-shuffled data.

Linear track task

A linear track (Figure 4.5a, 1.5 meters long and 200cm wide) was constructed with food receptacles at both ends. Wall mounted distal reference cues were kept in consistent positions for training and testing, as was track orientation in the room. This allowed food-ports (and hence run-directions) to maintain a consistent reward value across sessions. Rats ran between the two food-ports to retrieve a probabilistic reward, and three infrared beams (evenly positioned 25%, 50% and 75% along the track) provided an online assessment of rat position and run direction. A custom-written LABVIEW software monitored track beam-breaks and delivered rewards. The westward food-port was always associated with 100% reward probability (FP100), while eastward food-port was rewarded at 50% (FP50). Each trial started when rat breaks initial beam

(e.g. 25% beam near FP50 moving from FP50 -> FP100), causing it to prime for reward delivery. Crossing the 50% beam immediately delivered a reward at FP100 together with an audible reward tone (playback of finger snap, 70 dB). Crossing of the 75% beam completed the trial and initiated an ITI of 3 seconds to allow for reward collection and consumption. Other trials continued as described, with the exception of unrewarded trials where reward (and reward-cue) was omitted in half of FP50 trials.

For a more accurate analysis of rat position and trajectory offline, overhead color video was captured at 30 frames per second and stored to disk. Each frame was stamped with trial number and date/time (1ms precision). Rats were outfitted with very small (1.5 inch) green glow-stick (Amazon USA) to facilitate position tracking. A custom MATLAB routine was used to track rat position frame-by-frame as follows. First, video was cropped to the boundaries of linear track and we isolated video background by averaging all frames within a ~30 minute session. We then subtracted this background from each frame to find pixels that are changing (usually corresponding to rat). Lastly, we thresholded the green channel of camera output to isolate intense green signal from head-mounted glow-stick, and tracked the x and y position of thresholded centroid across all frames. Instantaneous velocity and acceleration were computed by performing a first or second order derivative (respectively) of x-position data with respect to time. To help streamline quantitative correlations between photometry data (sampled at 250 Hz) and positional changes (and their derivatives; i.e velocity and acceleration) obtained from video tracking (sample at 30 Hz), we re-sampled rat-position coordinates to match photometry sampling frequency. Because all sampled data points are stamped to a 1ms precision, we re-aligned positional data to neural samples, ensuring correspondence between each timestamp. Finally, we used linear interpolation to populate intervening missing data points.

Trial-and-error task

The trial-and-error task used in this chapter was identical to task described in chapter 2 and 3, with some modifications. Briefly, each trial started with the illumination of only the central port, awaiting instrumental responding. Rats must poke and maintain

the center poke for a variable hold duration of 0.5-1.5 seconds, followed by a 250ms long white noise 'GO' cue. Rats subsequently made rapid free choices to the rightward or leftward adjacent ports. If trial is rewarded, a sugar pellet is dispensed coincident with rat breaking side-port beam, generating an audible reward cue ('click' from food hopper). In unrewarded trials, breaking the side port beam omitted reward without any audio-visual cue. Exiting out of the side port demarcated the end of a trial, and a random ITI (6-10 seconds) was initiated. Entry into non-illuminated ports to start a trial (wrong starts) or violation of hold-duration (premature movement before 'GO' cue, classified as false starts) caused the house-light to turn on, aborting trial and initiating ITI. To promote cued responding, center-port illumination at beginning of a trial was delayed by additional 0.5 seconds if a response was detected at any of 5 ports upon ITI expiration. Left and right choices had independent reward probabilities, each maintained for blocks of 35-45 trials (randomly selected block length and sequence for each session). All combinations of 10%, 50% and 90% reward probability were used including 10:10 and 90:90.

Surgery

To achieve cell-type specific expression of the calcium indicator GCAMP6f (Chen et al., 2013) selectively in dopaminergic cells, we combined *TH-cre+* rats with CRE dependent viral constructs. Under isoflurane anesthesia, 15 rats received stereotaxic injection of AAV5-syn-FLEX-GCAMP6f (University of Pennsylvania vector core, 1.4 x 10^{13} viral particles ml^{-1}) bilaterally into the VTA (coordinates from bregma: -5.1mm and -5.4mm AP, 1.0mm ML, 7.0mm and 7.5mm DV from dura) for a total of 1.5uL per hemisphere. Pulled glass micropipettes pressure-ejected virus into the midbrain at 50 nL/min and left in place for 5 minutes to allow diffusion. After injection, the scalp was sutured shut, and rats were allowed to recover.

Three weeks after viral infusion, two rats received midbrain (coordinates: -5.2mm AP, 1.0mm ML) implant of optic ferrule (400µm core diameter, 0.49 NA) fused with a bipolar electrical stimulator (Figure 4.1b,c) under FSCV guidance. This opto-electrical stimulator was used to electrically stimulate DA cells, while simultaneously assaying

local GCAMP6f signal photometrically (Figure 4.1f,g). The exposed electrical contacts flanked the optic fiber, and terminated ~100µm ventral to optic fiber end (Figure 4.1c, inset). Under Ketamine/xylazine anesthesia, we lowed a carbon fiber electrode into the nucleus accumbens core (coordinates: 1.3mm AP, 1.3mm ML, 6.0mm DV from dura) while the opto-electric stimulator was lowered progressively to the VTA. We cemented the stimulator at a depth that maximized electrically evoked accumbens DA release, positioning the optic fiber above DA cells that directly project to the ventral stratum. Following 3 weeks of viral expression, the remaining 13 rats received unilateral stereotaxic implantation of optic ferrules into the right midbrain (targeting the VTA, coordinates same as above) and the right nucleus accumbens core (coordinates: 1.3mm AP, 1.3mm ML, 6.0mm DV from dura). Three of the fifteen rats did not exhibit viral expression at all, and, omitted from study. Midbrain fiber placement in 4 of remaining 12 rats was not found to target a population of DA cells that robustly exhibited GCAMP6 expression.

Photometry Instrumentation and Data Acquisition

Fiber photometry data was performed as previously described (Gunaydin et al., 2010; Lerner et al., 2015; Zalocusky et al., 2016) to assess the bulk changes in fluorescence from the midbrain. In principle, blue light illuminated GCAMP6f expressed in neural tissue, and upon binding of calcium, the green fluophore undergoes a conformational change to emit light in the green wavelength (Chen et al., 2013). Bulk fluorescence from hundreds (or thousands) of neurons is collected by an implanted optic fiber and focused on a very sensitive photodetector (or photon counter).

Under freely moving conditions, the small amount of emitted fluorescence is prone to movement-induced contamination (likely due to patch cable bending and inefficiencies associated with fiber coupling). One solution for this movement artifact is to use multicolor excitation in fast sequence. Specifically, 470nm blue light excitation of GCAMPis supplemented by 405nm violet light. This method exploits the isosbestic point of GCAMP6f (405nm, defined as a range of the excitation spectrum where GCAMP fluorescence is calcium invariable) to acquire a reference signal (Tian et al., 2009;

Tantama et al., 2012; Rizzuto and Szabadkai, 2014). If tuned appropriately, this reference signal is affected by movement artifacts to a similar extent (as 470 nm calcium dependent signal), and exhibits similar bleach time course.

An overall schematic of the photometry rig is presented in Figure 4.1a. Blue and violet light from SMA coupled 470 nm and 405 nm LEDs (each powered at 30uW) were first narrow-band-pass filtered (470/10 nm or 405/10 nm single band-pass filter, Semrock, Rochester NY) and combined using beam splitter cube (CM1-PBS, ThorLabs). The combined beam was then reflected off a dichroic mirror (DMLP505, Thorlabs NJ), and collimated (PAF-SMA-11-A, Thorlabs NJ) into a 3-meter long multimode SMA patch cable (400µm core UMT with SMA ferrule termination, Figure 4.1a). An optic ferrule implanted into rat VTA mated with the SMA patch cable (via zirconia sleeve), and delivered blue and violet light within a few hundred micrometers of GCAMP6f expressing cells (Figure 4.1b). Green fluorescence returned to the launchpad via 3 meter long SMA patch cable and through a dichroic mirror (Figure 4.1a). Another narrow-band-pass filter (520/10 nm single band-pass filter, Semrock, Rochester NY) cleaned up the returning signal before focusing onto an ultra-sensitive femtowatt detector (NewFocus 2151, Newport NJ) using an aspheric lens (352610-A, ThorLabs). The photodetector was set to DC mode (750 Hz bandwidth) and responded with a gain of 2 x 10^{10} Volts/Amp with a typical responsively of 0.4 Amp/Watt. Voltage output of photodetector was routed through lock-in amplifier (see below) before it was digitized.

To accomplish sequential sampling of GCAMP6f response to 470 and 405 excitation, we used carrier frequencies of 211Hz and 531Hz to modulate LED outputs. These frequencies were selected to minimize crosstalk-between the two channels and avoid contamination from 60 Hz overhead lights (Zalocusky et al., 2016), while remaining below the 750 Hz bandwidth of the photodetector. Two lock-in amplifiers (SR810, Standord Research systems) each dedicated to one channel, generated modulation frequencies to drive the LEDs and demodulated photodetector output according to lock-in frequency (settings: 500 V/µA sensitivity, 3ms time constant, 24 dB/ octave roll off, low noise reserve). A lock-in amplifier, in principle, functions to extract

small amplitude signals hidden in large amplitude broadband signal according to its carrier frequency, as is typically used in AM/FM radio. Here, we used the lock-in amplifier to extract the time course of photodetector signals related to either 470 or 405 light excitation according to their respective carrier frequencies. Demodulated signals were scaled to +/-10 V (to minimize digitization error) with a 16-bit ADC converter and stored to disk.

Data analysis

All data preprocessing and analysis made use of custom MATLAB routines. Photometry data (both 470 and 405 signals), operant box (or linear track) port states and trial numbers were all sampled simultaneously at 250Hz. We removed movement related artifacts and bleach time-course from 470 signal to retrieve a ratio-metric calcium dependent signal as described below. First, we used a linear least-squared fit to scale the 405 signal to the 470 as: fitted_405 = offset + a*470_sig. Next, this fitted signal was subtracted from the 470 and resulting transients were scaled according to extent of bleaching (delta_f/F = (470_signal - fitted_405) / fitted_405. delta_f/F was smoothed by a factor of 10 to achieve a single that varied at 25 Hz well within the bandwidth for GCAMP6f (~30Hz). Timestamps for various behavioral events were extracted from port states (also sampled at 250 Hz), and delta_f/F was aligned according to conditions of analysis (i.e. alignment time windows, or separation by different trial types).

Results

Midbrain photometric measurements

GCAMP6f signals from the VTA exhibited robust spontaneous activity, particularly during free locomotion, and reward consumption (Figure 4.1e). In two rats that received an electrical stimulator, we verified that photometric deflections were indeed neural signals, as brief electrical stimulations varying in train frequency, or pulse number reliably evoked signals with variable amplitudes and/or timecourse (Figure 4.1f,g).

Probabilistically rewarded pavlovian approach

Rats responded to CS tones that predicted reward at 0, 0.5 and 1.0 probability by displaying conditioned approach behaviors to the food receptacle (Figure 4.2b). To assess the conditioned behavioral expression of learned CS value, we first quantified approach latency defined as the earliest interruption of IR-beam at the food-port following CS onset. Response latency scaled according to reward expectation (Figure 4.2b), with rats approaching the food receptacle with haste in 1.0 probability trials, and displaying progressively longer latency for 0.5 and 0 probability conditions. A one-way ANOVA revealed a significant effect of reward probability on approach latency from the seven animals tested (Figure 4.2c, $F(2,12)=4.76$, $p = 0.03$). To further assess the time-varying probability of conditioned responding, we quantified the moment-by-moment IR-beam breaks aligned to CS onset (Figure 4.2d). Rats' probability of displaying conditioned responding to the food receptacle varied according to probability of reward. To test if rats successfully classify auditory tones based on their reward value, we performed area under receiver operator curve (auROC) analysis for responding during CS. auROC was significant in all 7 rats (mean auROC = 0.62, all p-values < 0.0001 for 100% CS and mean auROC = 0.58, all p-values < 0.05 for 50% CS) comparing anticipatory food-port entries during CS against baseline responding. Conditioned responding for 0% CS was not significantly different from baseline in all rats (mean auROC = 0.49, $p > 0.05$). Taken together, these results indicate that rats successfully learned the predictive value of the conditioned stimuli.

We next monitored bulk activation VTA during conditioning task. Consistent with prior reports (Oyama et al., 2010; Eshel et al., 2015) of midbrain DA activity during pavlovian conditioning, we observed robust phasic activation of the VTA at onset of CS tones and delivery of reward (Figure 4.3). Unpredicted rewards recruited DA cells with very short latency, reaching peak fluorescence within 200-400 ms(Figure 4.3a,b). This robust reward response was apparent in all rewarded trials, albeit varied according to reward expectation. Furthermore, CS tones also produced VTA responses with short latency which were also varied according to value. But, CS responses were smaller than US reward responses, likely due to the short duration of training in our task (2-3 weeks, total of ~500 trials per rat), unlike extensive (thousands of trials) training in monkeys (Kobayashi and Schultz, 2008; Fiorillo et al., 2008; Mirenowicz and Schultz, 1994).

To assess if the onset of CS modulated VTA activity (Figure 4.3d), we performed a one-way ANOVA with the four trial types as levels of the factor 'EXPECTATION'. We found that VTA activity following CS onset (peak level during one second after CS onset) was significantly affected by Expectation ($F(3,4) = 22.5$, $p = 0.006$). But a post hoc test for differences among the group revealed that CS responses were significantly elevated relative to the same epoch in unexpected reward condition (all $p < 0.01$), but did not significantly differ among each other (all p-values greater than 0.3). In other words, the CS produced increased activity in VTA irrespective reward expectation signaled by tone. We next assed if VTA activity during the outcome epoch exhibited variability depending on tone-specific reward expectations. One-way ANOVA suggested that there were no significant differences between VTA activity during reward epoch (EXPECTATION main effect $F(2,5) = 2.88$, $p = 0.15$). Together these results suggest that, in our dataset, VTA activity is robustly modulated by cues associated with rewards, but do not signal the value of cues, or errors associated with delivery of rewards. It should be noted that the individual session and aggregate data (see figure 4.3b,c) are certainly trending in the direction of previous reports of DA coding of CS value and prediction errors. Our lack of achieving statistical significance is potentially due to under training and variability of our signals from rat to rat.

Linear track

Well-trained rats ran back-and-forth the length of track to retrieve rewards (Figure 4.5b), performing 195 trials +/- 51 SEM during each 30-minute session. We used overheard video tracking of rat position to quantify changes in displacement for run trajectory toward either the 100% direction (to FP100, Figure 4.5b) or 50% port direction (to FP50, Figure 4.5c). We further transformed displacement into velocity and acceleration as an index of vigorous pursuit of rewards. To assess how trial-by-trial vigor is affected by reward expectation, we separated instantaneous velocity and acceleration according to trial types. Rats progressively ramped their velocity and reached significantly higher pre-outcome velocity when running toward the higher value (FP100) compared to starting movement toward FP50 (two tail t-test comparing mean pre outcome velocity for FP100 and FP50, $p = 0.042$). In addition, following the reward cue in FP50 trials, rats were additionally invigorated as they ramped their velocity to retrieve reward (two tail t-test comparing mean post outcome velocity for rewarded and unrewarded trials, $p < 0.001$). This distinction was apparent when we aligned velocity both according to time-to-reward-cue, and position on the track (Figure 4.5a,b,e). Trajectories toward FP100 were associated with brief bursts of acceleration that peaked sooner, and reached higher levels relative to FP50 trials (Figure 4.5c,d,f). Together, these results suggest that rats learned the respective reward-value of each food port and adapted their initial performance-vigor (velocity) according to reward expectation, and transiently modified their motivation (acceleration) following surprising events. In other words, this task selectively varied trial-by-trial reward expectation and prediction errors, and rats readily modify their behaviors.

To distinguish the time course of brain signals for reward expectation and prediction error, we first considered the dynamics of idealized *RPE* and *Value* signals under the conditions of this task. Overall, *Value* signals are expected to ramp as rats get closer to the goal in a manner consistent with temporal discounting (see chapter 2), whereas *RPE* signals should transiently peak during the feedback epoch (Figure 4.6a). Pre-outcome *Value* signal for FP100 will be higher in comparison to FP50 movement

trajectory, but *error* signals will be zero for both trial types during this epoch. Once outcome is revealed, *Value* signals in rewarded trials both under 100% and 50% conditions will converge and continue to increase until rats reach goal location(Figure 4.6a, top). *RPE* signals, on the other hand, will be transiently activated during the outcome epoch, scaling according to the degree of surprise associated with reward delivery or omission (Figure 4.6a, bottom). Given these temporal dynamics of *Value* and *RPE* signals, we make the following three specific predictions: If a signal codes for a *Value* signal, 1) there will be statistically different levels of activation during the pre-outcome epoch among FP100 and FP50 trials (Figure 4.6b, left). 2) Post-outcome levels are predicted to be equivalent for all rewarded trials, but drops to zero for unrewarded trials (Figure 4.6b, middle). 3) All rewarded trials will achieve maximal levels when the rat is at the goal (Figure 4.6b, right). On the other hand, if a signal codes for *RPE*, 1) Pre-outcome epoch levels will be indistinguishable in the two trial types, or from zero (Figure 4.6c, left). 2) Post-outcome levels will scale according to degree of surprise (Figure 4.6c, middle), and 3) Peak *RPE* signals are observed briefly during the outcome epoch, and not when the animal is at goal location (Figure 4.6c, right).

We observed that rewarded trials in both FP100 and FP50 evoked brief activation of the VTA selectively during the outcome epoch(Figure 4.6d). Additionally, unrewarded trials were accompanied by transient dips in VTA activity that sometimes fell below baseline. Figure 4.7 shows the time course of activity for each rat. To test our hypothesis that VTA signals quantitatively signal *RPE*, and not a *Value* signal, we analyzed photometry data according to predictions.

First, VTA responses during the pre-outcome epoch were not different under 50% and 100% reward expectation (Figure 4.6e left, two tail t-test comparing mean post outcome DA for FP100 and FP50, $p = 0.66$), and responses for both trial types were not significantly different to zero ($p = 0.4$ and $p = 0.76$ respectively for FP100 and FP50). Second, VTA activity during a 1-second epoch following reward cue scaled according to experienced *RPE* (Figure 4.6e, middle $r=0.85$), and did not match the prediction of a

Value account as reward receipt in the two trial types recruited the VTA to a significantly different extent (two tail t-test comparing mean post outcome DA for rewarded and unrewarded conditions, $p = 0.0001$). Finally, trial-by-trial peak VTA activation was observed just after rat crossing the half way location on the track(Figure 4.6e, right), producing distribution of peak position and time significantly different from goal location ($p = 0.021$, ttest). This finding is a direct contrast to previously reported voltammetry results where accumbens DA levels continuously increase, and reach peak levels when rat arrives at goal destination.

Trial-and-error task

Behavioral training on the trial-and-error task was performed as previously described in chapter 2 and rats adapted their behavior to changing reward contingencies in a similar manner. Notably, addition of 90:90 and 10:10 blocks did not qualitatively change adaptive vigor or choice behavior (data not shown).

VTA activity assayed via photometry displayed strong modulation during several behavioral events and trial types (Figure 4.8). Notably, VTA DA neurons responded to center port illumination abruptly with short latency and this response was apparent in all sessions (Figure 4.8). This was in contrast to our previous voltammetry report where early trial release was better aligned to the approach behavior than illumination of port light (Figure 2.3c). We further observed a modest increase in VTA activation accompanying the approach behavior, although not consistently observed in all session (Figure 4.9). During the hold period, the VTA maintained a modest plateau of responding, but longer wait durations produced a progressive decrease in calcium signal that dipped below baseline in some sessions. This observation is consistent with the notion that VTA cells are involved in signaling the temporal prediction of rewards, and the timing of cues that provide information about impending reward.

Much like responses observed for DA release, VTA activation also strongly distinguished rewarded and unrewarded trials when outcomes are revealed. The abrupt ‘click’ produced by food hopper delivering a sugar pellet caused an abrupt large

increase which peaked within 200-400ms of side-port entry. By contrast, DA release measured by voltammetry peaked 1-2 seconds after side port entry, demonstrating that VTA response considerably precede release within the striatum. On the other hand, omission of rewards during the outcome epoch produced a rapid decline in VTA activity that was observed to dip below pre-trial levels. To make a crude estimate of rat's reward expectation on any given trial, we counted the number of previously attained rewards in a rolling 10 trial window. We next correlated the strength of VTA activation (or suppression) during this outcome epoch to reward expectation. We observed a strong relationship between reward response in a 1 second time window and trial-by-trial reward expectation (average spearman's correlation coefficient = -0.34, p-value significant in 6 of 7 rats, range = [-0.58 0.05]). Conversely, unrewarded responses correlated with reward expectation to a weaker degree (average r = -0.13, significant in 3 of 7 sessions, range = [-0.29 -0.01]).

On rewarded trials, rats pull out of the side ports and approach the food receptacle to retrieve rewards. We have previously reported that in the majority of tested rats, approaching this goal destination yielded the largest intra-trial DA (Hamid et al., 2015). By contrast however, VTA cells did not ramp their response as rats approach food-port, but rather reactivated after rats enters the food receptacle (Figure 4.8 far right). It is noteworthy that VTA responses for rewarded and unrewarded outcomes were surprisingly transient in nature (lasting 0.5 to 1 second), in contrast to striatal DA concentrations that remained elevated (or depressed) for several seconds (Figure 2.3).

Discussion

In this study, I quantified bulk fluorescence of GCAMP dynamics from deep midbrain DA nuclei. I performed these measurements in rats performing three different behavioral tasks designed to assess brain signals for value and reward prediction errors. This chapter made some technical breakthroughs, but for a variety of reasons did not achieve its intended goal of directly contrasting midbrain DA signals to striatal activity. Nonetheless, I provide some preliminary conclusions below.

In all tasks performed, we observed strong fluctuations in VTA activity, but sometimes did not observe anticipated results. Specifically, in the pavlovian task, DA cells were clearly recruited by onset of CS and delivery of rewards, but across multiple rats, we did not observe significant variability of photometric signals depending on CS value. Indeed, midbrain DA cells have been extensively documented to be modulated by CS value and reward surprise (Ljungberg et al., 1992; Schultz et al., 1993; Oyama et al., 2010; Cohen et al., 2012b; Eshel et al., 2015; Eshel et al., 2016). We noted that our data trended in the direction of previous findings, but did not achieve statistical significance likely due to short training duration, and variability in photometry signals. This former suggestion is not unwarranted, previous groups have reported that maximal CS/US responses of DA cells generally requires substantial training (hundreds or thousands of trials) (Pan and Hyland, 2005; Kobayashi and Schultz, 2008; Redgrave et al., 2008).

The linear track task provides strong evidence for midbrain coding of reward prediction errors, and not reward value. It has been demonstrated before that accumbens DA release continuously ramps as rats approach food port on a maze task (Howe et al., 2013). These escalating DA ramps distinguished reward magnitude at goal locations, and were not transient in nature, gradually increasing, and signaling reward proximity and value (similar to out ‘value’ prediction in figure 4.6a). On the contrary, photometric VTA signals from our study appeared to signal RPE, transiently and selectively activating during the outcome epoch. Further, instead of correlating with value of anticipated rewards, we found that VTA activity scaled according to reward

surprise. Together with data from chapter 2 and previous linear track report (Howe et al., 2013), our linear track finding strongly suggests a disparate activity pattern of VTA and striatal DA release.

Lastly, I provided a first pass analysis of VTA activity in the trial-and-error task. DA cells were robustly modulated by several behavioral events in task. The contrast between early-trial VTA activity and NAcc DA release was noteworthy. In chapter 2 we noted that early trial increases accompanied approach behaviors, and were not locked to center port illumination. On the other hand, VTA cells are transiently activated by center port illumination, particularly if the rat was already engaged (i.e. short latency trials). Midbrain DA cells abruptly increased their activity in response to 'GO' cue. Interestingly, more delayed 'GO' cue appearance was associated with progressive decline in VTA calcium signals. This finding is consistent with previous report of DA cell coding of moment-by-moment hazard rate of reward related cues (Pasquereau and Turner, 2015). Finally, VTA activity during the outcome epoch clearly distinguished rewarded and unrewarded trials. Further, this outcome activity was strongly correlated with reward surprise.

In this chapter, I set out to study if DA signaling on the midbrain-forebrain axis was different, and to explore potential mechanisms underlying this divergence. Despite multiple triumphs in behavioral, technical, experimental and analysis fronts, some challenges preclude a clear conclusion for these experiments. While multiple rats displayed robust VTA signals, striatal photometry was not consistent across animals and did not yield high signal-to-noise ratio. One possibility may be that I only waited 5-8 weeks after virus injection, a timeframe likely insufficient for robust GCAMP trafficking into DA terminal fields.

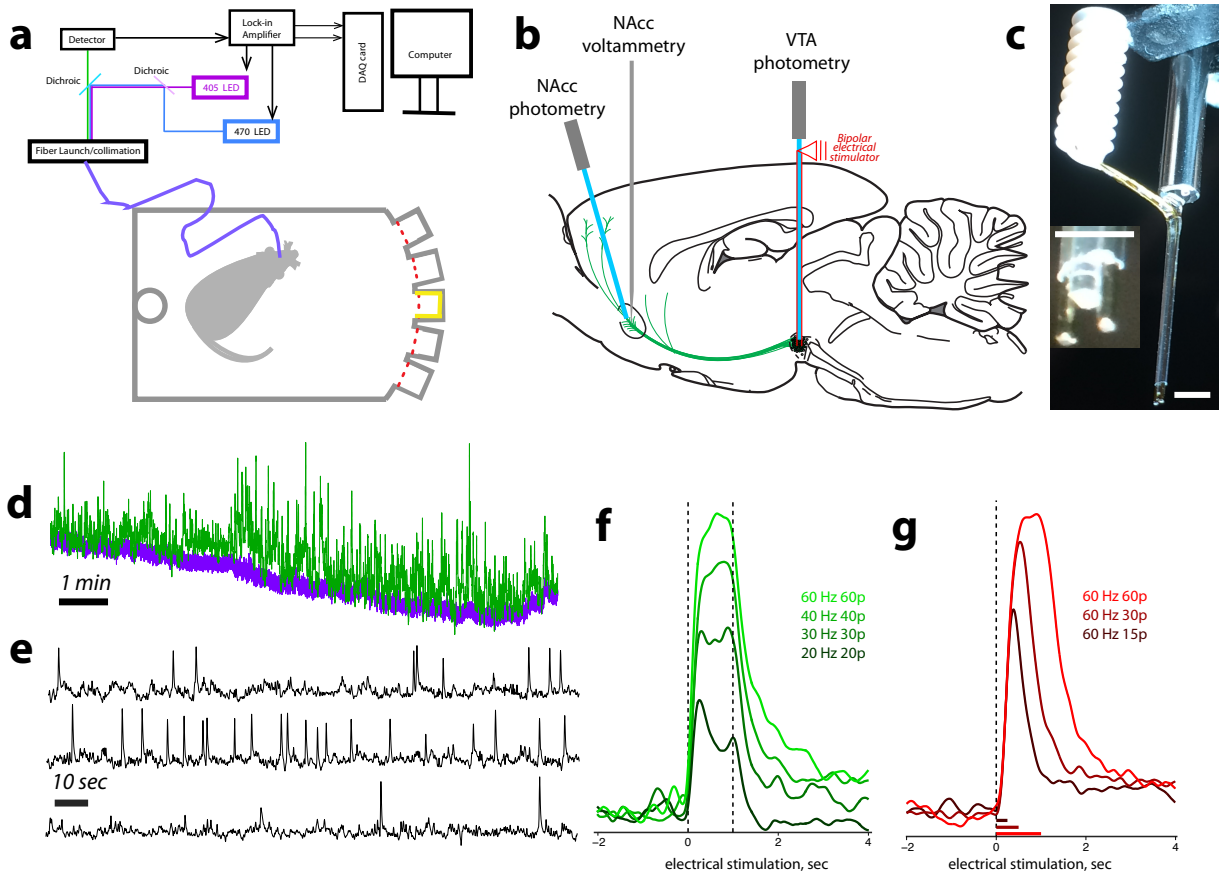

Figure 4.1: Photometry instrumentation and verification of fluorescent signals.

(a) Overview of key components of fiber photometry. A rat was connected to acquisition setup via a 3 meter long patch cable (purple). Two LED (blue and violet) diodes provided excitation, and returning green light was focused onto detector. Voltage output of the detector was routed through lock-in amplifier, and digitized by data acquisition card on a computer. (b) Schematic for implantation. Rats received GCAMP6f injection into the VTA. We cemented optic fiber together with bipolar electric stimulator into VTA under voltammetry guidance in NAcc, which also received optic fiber implant. (c) Close up photograph of an opto-electric stimulator. We glued an optic fiber to a bipolar stimulator, and exposed the electrical ends <100µm from optic fiber termination (inset). Scale bars represent 1mm in main image, and 100um in inset. (d) Raw data for 405 (purple trace) and 470 (green trace) modulated channels. Note overall bleach trend apparent in both signals, but green trace additionally contains transient changes

corresponding to neural activity. (e) Fiber photometry signals after data were processed (see methods) for removal of movement artifact and bleach time-course. Spontaneous VTA activity was observed as rat explored chamber. (f) To very that measured fluorescent signals were from VTA DA cells, we delivered electrical current (140μA) through the bipolar stimulator, while simultaneously monitoring changes in fluorescence. VTA fiber photometry responded to stimulations of different (f) frequency (constant pulse-train duration) and (g) different pulse durations (constant frequency).

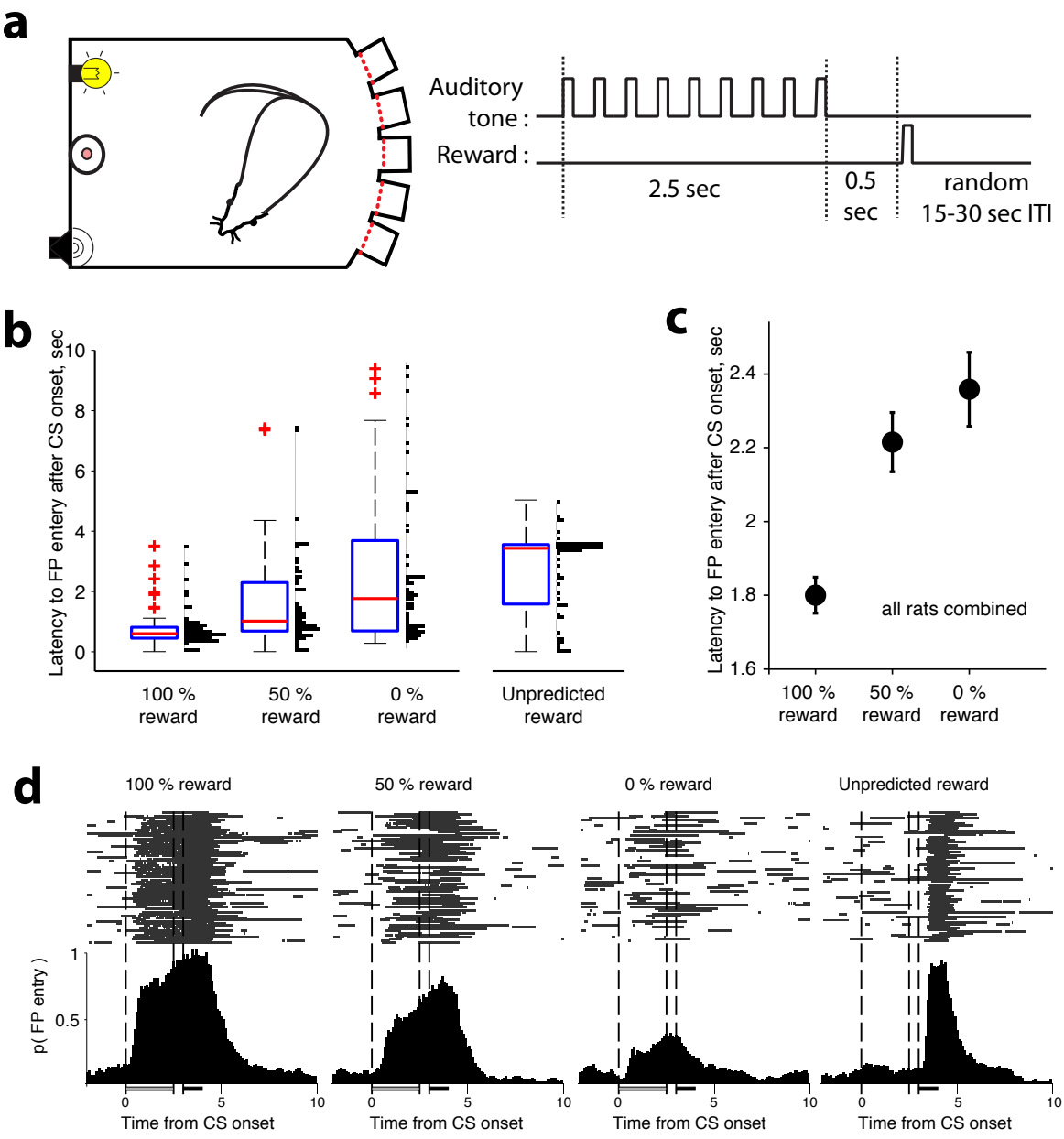

Figure 4.2: Pavlovian conditioned approach task and behavior.

(a) Schematic of operant chamber events of one trial. Rats performed task in the same box as the trial-and-error task, but overhead houselights was illuminated for duration of task. An auditory conditioned stimulus consisting many tone-bips was played for 2.5 seconds, and a reward was derived after a 0.5 second delay. (b) Box plots and distributions of latency to enter food port following CS tones signaling probabilistic

reward from a sample session. (c) Rats displayed progressively delayed latencies for tones that predicted less certain rewards. (d) Moment-by-moment food port beam-breaks (100ms bins) surrounding tones and food delivery in a representative session. Rats displayed anticipatory approach, and entry to food receptacle more strongly for highly valued rewards. Responding during the same epoch for unpredicted reward (far right) which is not preceded by CS was used as 'baseline' response rate.

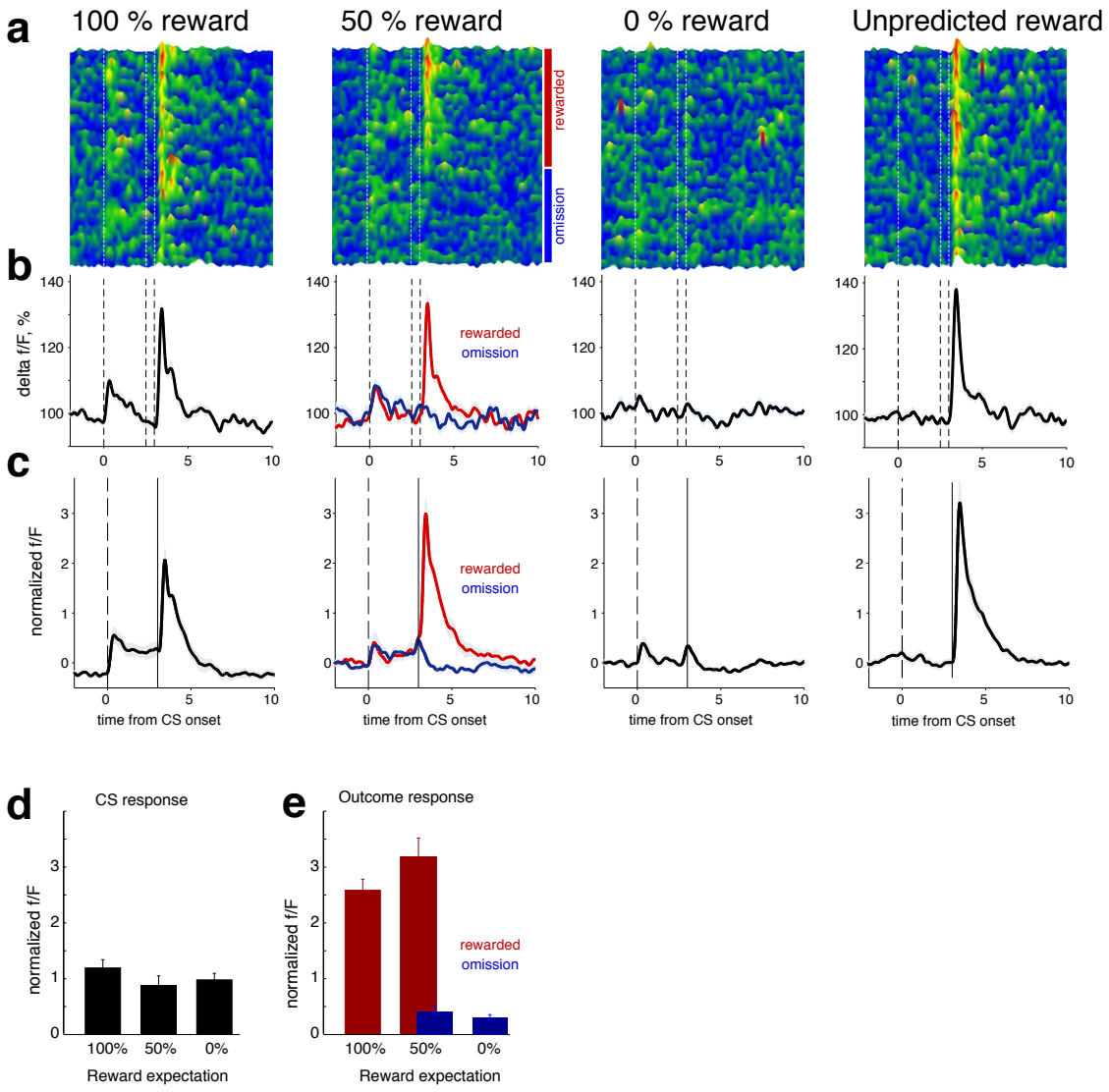

Figure 4.3: Activation of midbrain DA cells during pavlovian task performance.

(a)Trial-by-trial changes in VTA activity aligned to onset of CS tone, separated by tone signaling 100%, 50%, 0% or unexpected reward. (b) Average activity for the same conditions. Rewarded and unrewarded trials are separated for 50% condition. Data from one representative session. (c) Data from 7 rats was first peak normalized, and combined to assess the time course of VTA activity across all animals. We quantified the peak VTA activity following (d) CS onset or (e) reward delivery (both, one second epoch) for 7 rats.

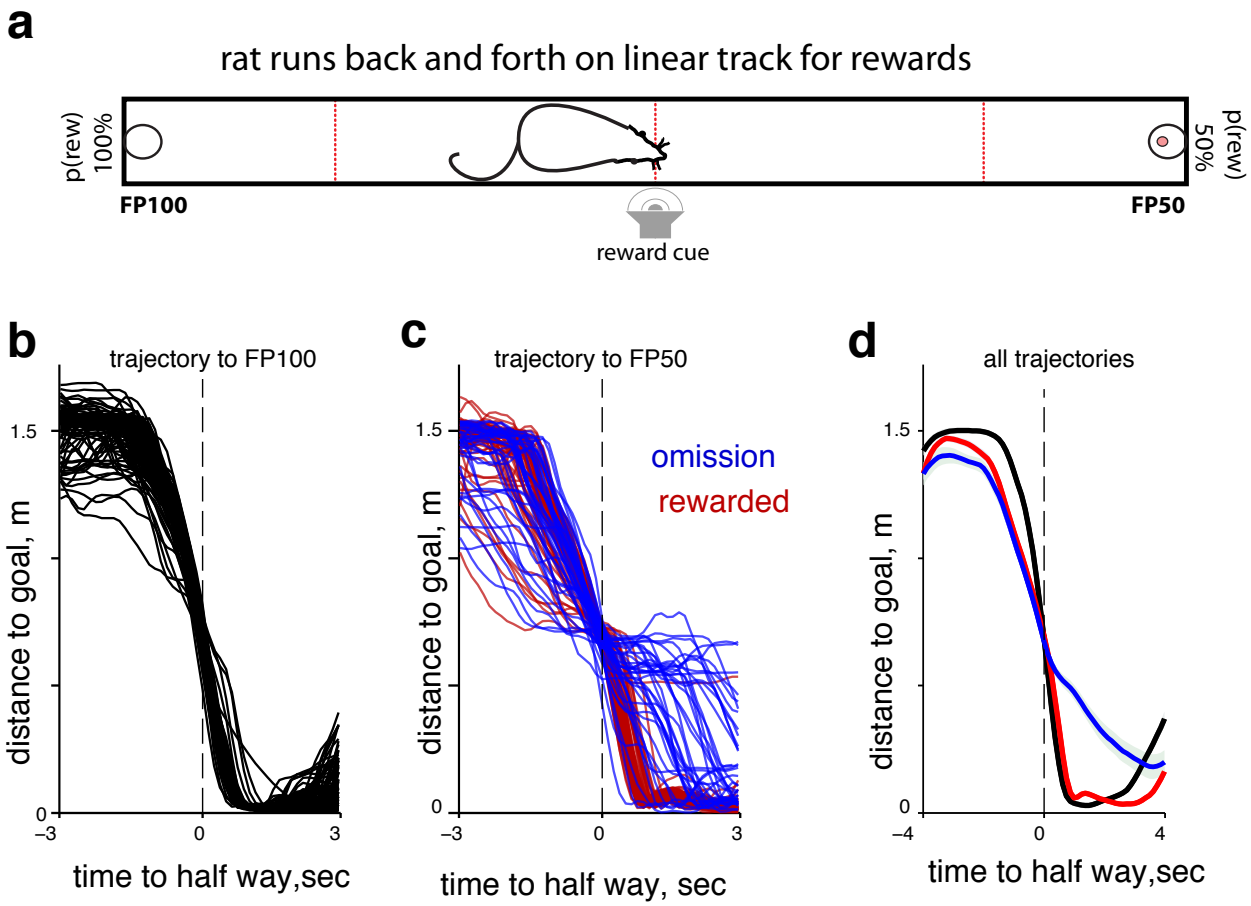

Figure 4.4: Behavioral performance on the linear Track.

(a) Linear track was constructed with food ports yielding rewards at 50% or 100%. On each trial, crossing of the middle beam delivered rewards, together with a reward cue (playback of finger snap). Overhead video was processed (see methods for details) to extract the moment-by-moment position of rat. We separated run trajectories for running toward (b) 100% food port (FP100) and (b) 50% food port (FP50, red for rewarded condition and blue for unrewarded condition). (c) Post outcome displacement trajectories are dependent on receipt of reward.

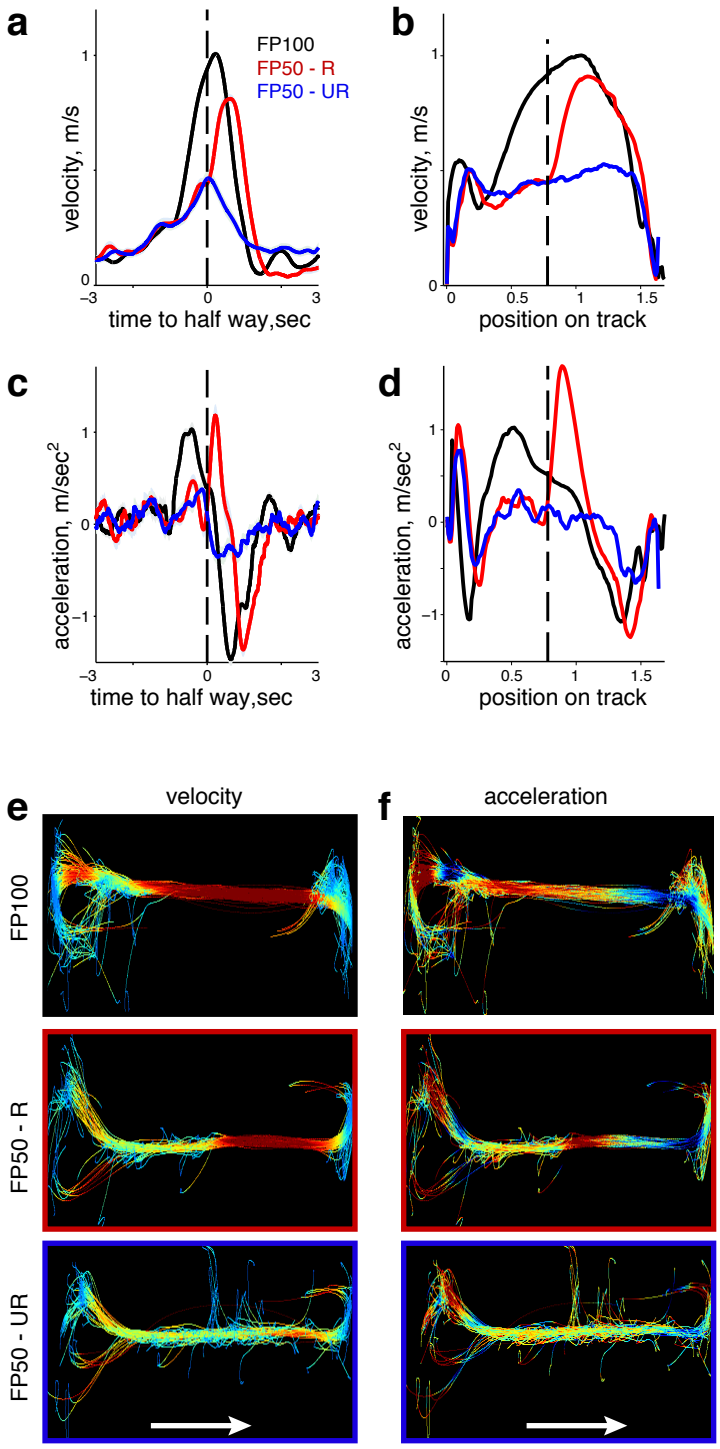

Figure 4.5: Changes in velocity and acceleration on linear track

Average time courses of velocity and acceleration changes as rat approaches the half way point on the linear track in a representative session. This half way point is

associated with the reward cue. Velocity changes to time (a) or space (b) are displayed, and changes in acceleration are shown in (c) and (d). (e) mean velocity broken down by run direction and reward plotted as a function of spatial location of the animal. Arrow at bottom shows the direction for all plots (i.e. goal food port is always shown on the right). (f) mean acceleration in the same format as (e).

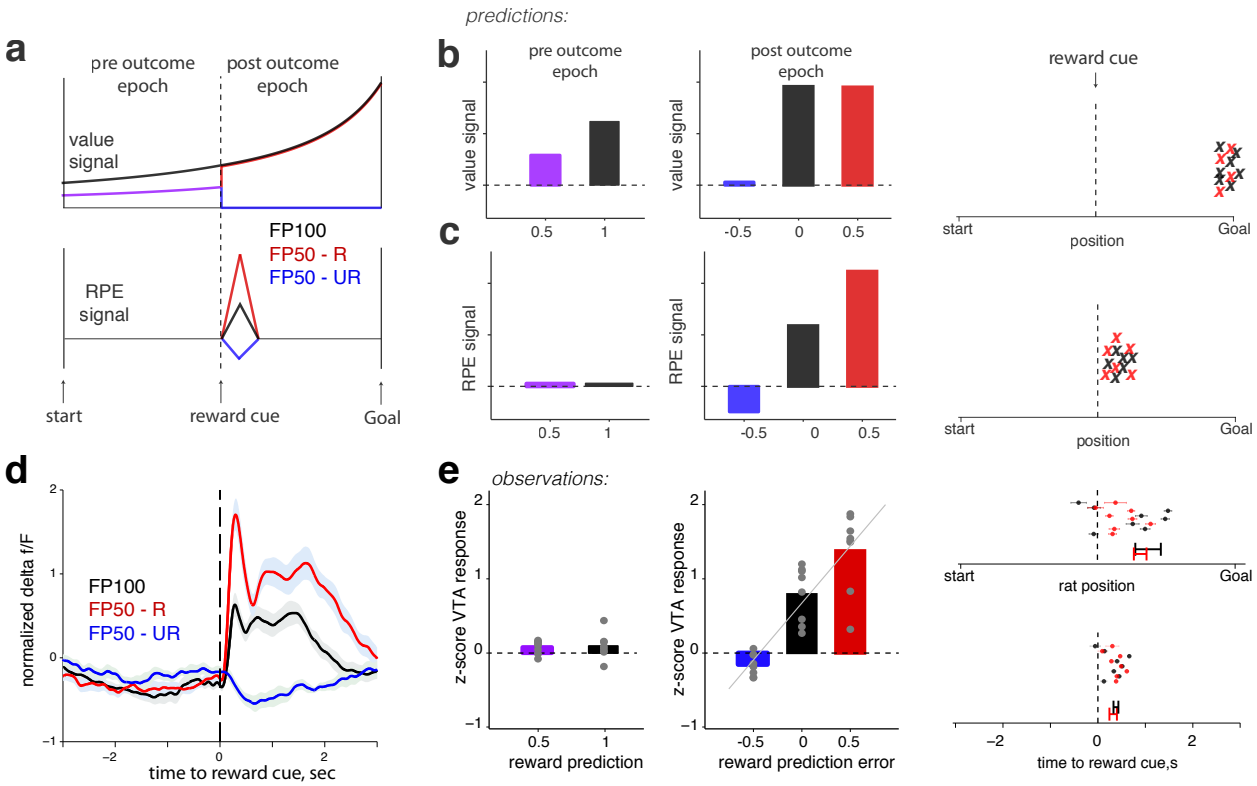

Figure 4.6: Testing prediction of VTA activity hypothesis.

(a) Top: predicted time course of a V*alue* signal on the linear track. Before outcome is revealed, the discounted value of trajectory for 100% reward (black) is higher than that of 50% reward (purple). Once outcome is signaled by the reward cue, *Value* for rewarded trials reflect the discounted value of the certain reward (black trace [100% probability, rewarded] and red trace [50% probability, rewarded]), while unrewarded trials fall to zero value for remainder of trajectory (blue). Bottom: prediction of changes in *RPE* during probabilistic linear task performance. Error signals are only apparent if new information is available from the environment. Hence, *RPE* signals are robustly active only during feedback epoch. Following the reward cue, RPE reflects the degree of surprise associated with reward. Rewarded trials for 100% trajectory mildly activates RPE signals (black), whereas rewards for 50% trials recruits large error signals (red). (b) Quantitative predictions of *Value* signals. Left: during the pre-outcome epoch, value for 50% (purple) and 100% (black) trajectories are different. Middle: *Value* during the

post-outcome epoch are identical for rewarded trials, but zero for unrewarded trials. Right: we predict that value reaches peak levels at the goal location. (c) Quantitative predictions for *RPE* time course, same format as (b). Left, pre outcome RPE for both trial types is zero. middle: there is a linear scaling of RPE signals depending surprise. and RPE signals are expected to reach peak levels just after the reward cue (or having crossed the mid-way location on track). (d) Peak normalize average VTA photometry for all rats aligned to reward cue (n=6 rats). Data was first averaged for each rat (195 trials +/- 51 SEM per rat). (e) Left: quantification of pre-outcome (defined as one second duration before reward cue) VTA activity for 50% and 100% trajectory. Each grey dot represents average from each rat. Mean levels for the two conditions were not different from each other (p=0.66, ttest), or from zero(ttest, p=0.4 and p=0.77 respectively for 100% and 50% trajectory). Middle: VTA activity during the outcome epoch (one second after reward cue) scaled according to reward surprise (n = 6 rats. spearman r = 0.85, $p<0.001$, one way ANOVA with expectation as factor $F(2,4) = 155.6$, $p = 0.00017$, post hoc comparison revealing a significant difference between all three conditions). Right: Peak activity VTA activation level was observed just after reward cue, and significantly different from goal location ($p = 0.021$, ttest).

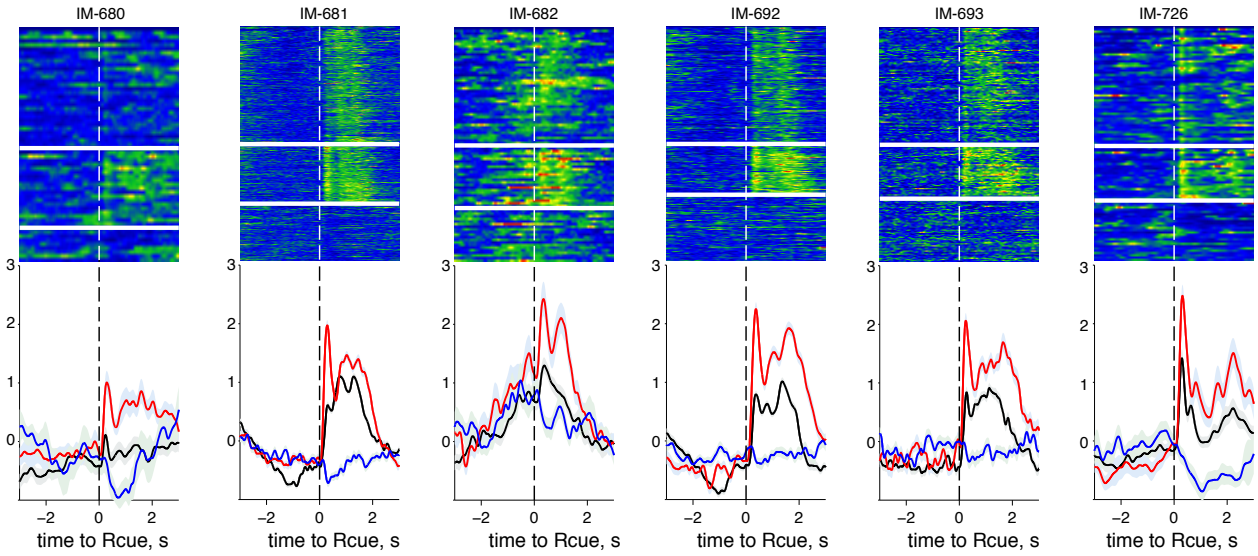

Figure 4.7: Individual session breakdown of linear track DA

VTA activity for 6 rat that was combined for 4.6d with rat identification at top. Rats performed different number of trials during 30 minute session. Rats completed 60,426,91,265,227,104 trials in the order they are displayed left to right.

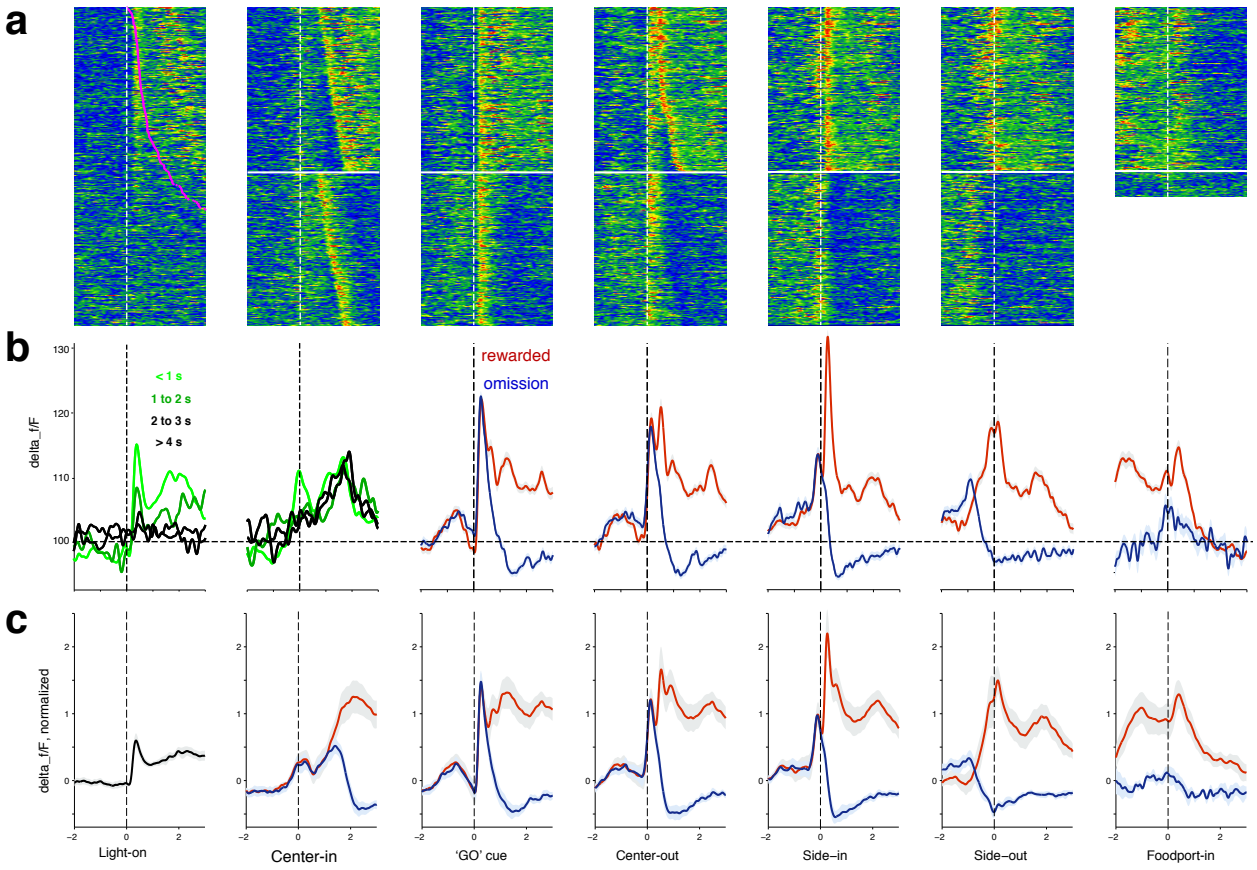

Figure 4.8: VTA dynamics during RL task performance

(a) VTA photometry during a representative rat performance of trial-and-error task. Each row of color plot represents a single trial. Trials are sorted by behavioral event following alignment. Solid horizontal line separate rewarded and unrewarded trials. (b) Mean time course for VTA signals. Light-on and Center-in alignments are broken down by latency and other alignments are broken down by rewarded/unrewarded. (c) Data for each animals was averaged, and aggregate data from across all sessions are displayed.

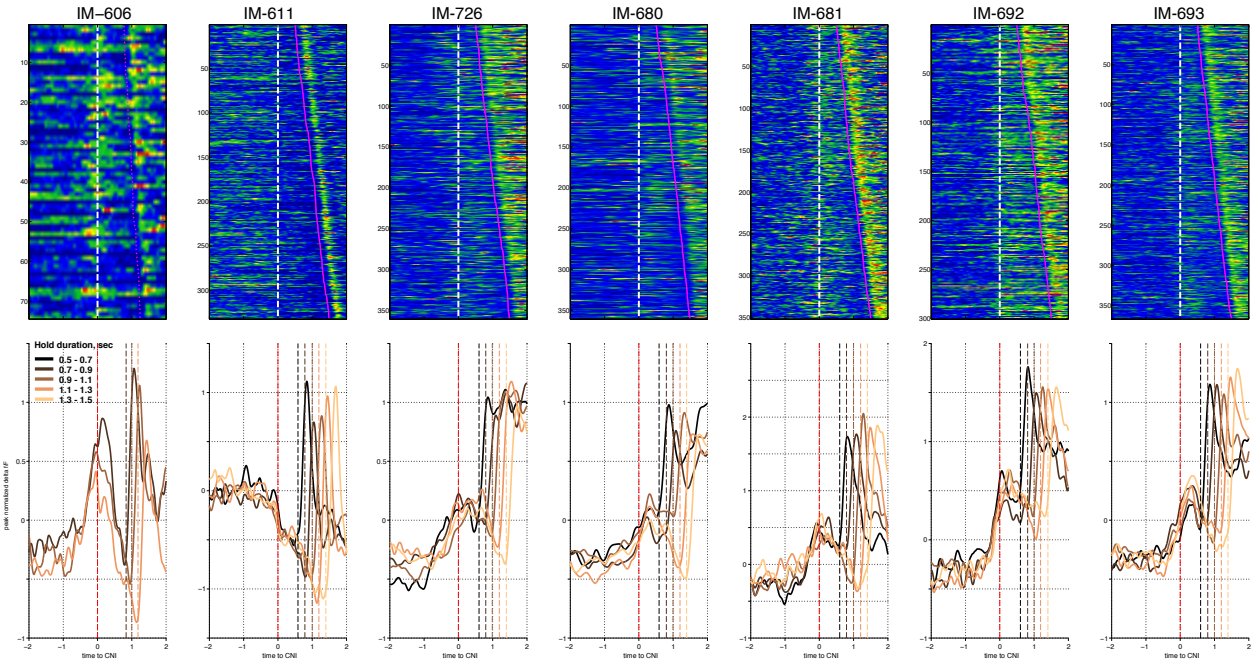

Figure 4.9: Individual session signaling of time to cue.

VTA activity was aligned to Center-in event and sorted based on the hold duration before GO. Top: each row of color plot shows a single trial for each rat (rat identification at top). White line is Center-in and pink dots mark the appearance of 'GO' cue for each trial. Bottom: we broke down each mean VTA activity by delay to cue (in 200ms bins) aligned to Center-in.With the exception of one rat (IM-611) note the overall ramp and plateau of VTA activity upon center entry. Delayed appearance of GO cue however is associated with a progressively decreasing level.

References

Akerboom, J., Chen, T.-W., Wardill, T.J., Tian, L., Marvin, J.S., Mutlu, S., Calderón, N.C., Esposti, F., Borghuis, B.G., Sun, X.R., Gordus, A., Orger, M.B., Portugues, R., Engert, F., Macklin, J.J., Filosa, A., Aggarwal, A., Kerr, R.A., Takagi, R., Kracun, S., Shigetomi, E., Khakh, B.S., Baier, H., Lagnado, L., Wang, S.S.-H., Bargmann, C.I., Kimmel, B.E., Jayaraman, V., Svoboda, K., Kim, D.S., Schreiter, E.R., and Looger, L.L. (2012). Optimization of a GCaMP Calcium Indicator for Neural Activity Imaging. J Neurosci *32*, 13819-13840.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129-141.

Chen, T.W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., Looger, L.L., Svoboda, K., and Kim, D.S. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature *499*, 295-300.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012a). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85-88.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012b). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85-88.

Cragg, S.J. (2006). Meaningful silences: how dopamine listens to the ACh pause. Trends Neurosci *29*, 125-131.

Cui, G., Jun, S.B., Jin, X., Pham, M.D., Vogel, S.S., Lovinger, D.M., and Costa, R.M. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. Nature *494*, 238-242.

Day, J.J., Roitman, M.F., Wightman, R.M., and Carelli, R.M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci *10*, 1020-1028.

Denk, W., Strickler, J.H., and Webb, W.W. (1990). Two-photon laser scanning fluorescence microscopy. Science *248*, 73-76.

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. Nature

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. Nat Neurosci *19*, 479-486.

Fiorillo, C.D., Newsome, W.T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. Nat Neurosci *11*, 966-973.

Flagel, S.B., Clark, J.J., Robinson, T.E., Mayo, L., Czuj, A., Willuhn, I., Akers, C.A., Clinton, S.M., Phillips, P.E., and Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. Nature *469*, 53-57.

Flusberg, B.A., Nimmerjahn, A., Cocker, E.D., Mukamel, E.A., Barretto, R.P.J., Ko, T.H., Burns, L.D., Jung, J.C., and Schnitzer, M.J. (2008). High-speed, miniaturized fluorescence microscopy in freely moving mice. Nat Methods *5*, 935-938.

Gage, G.J., Stoetzner, C.R., Wiltschko, A.B., and Berke, J.D. (2010). Selective activation of striatal fast-spiking interneurons during choice execution. Neuron *67*, 466-479.

García, A.G., García-De-Diego, A.M., Gandía, L., Borges, R., and García-Sancho, J. (2006). Calcium signaling and exocytosis in adrenal chromaffin cells. Physiol Rev *86*, 1093-1131.

Glimcher, P.W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A *108 Suppl 3*, 15647-15654.

Gunaydin, L., Grosenick, L., Finkelstein, J., Kauvar, I., Fenno, L., Adhikari, A., Lammel, S., Mirzabekov, J., Airan, R., Zalocusky, K., Tye, K., Anikeeva, P., Malenka, R., and Deisseroth, K. (2014). Natural Neural Projection Dynamics Underlying Social Behavior. Cell *157*, 1535-1551.

Gunaydin, L.A., Yizhar, O., Berndt, A., Sohal, V.S., Deisseroth, K., and Hegemann, P. (2010). Ultrafast optogenetic control. Nat Neurosci *13*, 387-392.

Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2015). Mesolimbic dopamine signals the value of work. Nat Neurosci

Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. Nat Neurosci *19*, 117-126.

Helmchen, F., Fee, M.S., Tank, D.W., and Denk, W. (2001). A Miniature Head-Mounted Two-Photon Microscope. Neuron *31*, 903-912.

Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E., and Graybiel, A.M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature

Ikemoto, S. (2007). Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. Brain Res Rev *56*, 27-78.

Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. J Neurosci *28*, 7837-7846.

Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B., and Roeper, J. (2008). Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. Neuron *57*, 760-773.

Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K., Malenka, R.C., Luo, L., Tomer, R., and Deisseroth, K. (2015). Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell *162*, 635-647.

Leventhal, D.K., Gage, G.J., Schmidt, R., Pettibone, J.R., Case, A.C., and Berke, J.D. (2012). Basal ganglia beta oscillations accompany cue utilization. Neuron *73*, 523-536.

Leventhal, D.K., Stoetzner, C.R., Abraham, R., Pettibone, J., DeMarco, K., and Berke, J.D. (2014). Dissociable effects of dopamine on learning and performance within sensorimotor striatum. Basal Ganglia *4*, 43-54.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. J Neurophysiol *67*, 145-163.

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837-841.

Mirenowicz, J., and Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. J Neurophysiol *72*, 1024-1027.

Onge, J.R.S., Ahn, S., Phillips, A.G., and Floresco, S.B. (2012). Dynamic fluctuations in dopamine efflux in the prefrontal cortex and nucleus accumbens during risk-based decision making. The Journal of Neuroscience *32*, 16880-16891.

Oyama, K., Hernádi, I., Iijima, T., and Tsutsui, K. (2010). Reward prediction error coding in dorsal striatal neurons. J Neurosci *30*, 11447-11457.

Pan, W.X., and Hyland, B.I. (2005). Pedunculopontine tegmental nucleus controls conditioned responses of midbrain dopamine neurons in behaving rats. J Neurosci *25*, 4725-4732.

Parker, N.F., Cameron, C.M., Taliaferro, J.P., Lee, J., Choi, J.Y., Davidson, T.J., Daw, N.D., and Witten, I.B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. Nat Neurosci

Pasquereau, B., and Turner, R.S. (2015). Dopamine neurons encode errors in predicting movement trigger occurrence. J Neurophysiol *113*, 1110-1123.

Pereira, D.B., Schmitz, Y., Mészáros, J., Merchant, P., Hu, G., Li, S., Henke, A., Lizardi-Ortiz, J.E., Karpowicz, R.J., Morgenstern, T.J., Sonders, M.S., Kanter, E., Rodriguez, P.C., Mosharov, E.V., Sames, D., and Sulzer, D. (2016). Fluorescent false neurotransmitter reveals functionally silent dopamine vesicle clusters in the striatum. Nat Neurosci

Petersen, O.H., and Ueda, N. (1976). Pancreatic acinar cells: the role of calcium in stimulus-secretion coupling. J Physiol *254*, 583-606.

Phillips, P.E., Stuber, G.D., Heien, M.L., Wightman, R.M., and Carelli, R.M. (2003). Subsecond dopamine release promotes cocaine seeking. Nature *422*, 614-618.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? Brain Res Rev *58*, 322-339.

Rice, M.E., and Cragg, S.J. (2004). Nicotine amplifies reward-related dopamine signals in striatum. Nat Neurosci *7*, 583-584.

Rizzuto, R., and Szabadkai, G. (2014). Measuring baseline Ca2+ levels in subcellular compartments using genetically engineered fluorescent indicators. Cell-wide Metabolic Alterations Associated With Malignancy *543*, 47.

Roesch, M.R., Calu, D.J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci *10*, 1615-1624.

Roitman, M.F., Stuber, G.D., Phillips, P.E., Wightman, R.M., and Carelli, R.M. (2004). Dopamine operates as a subsecond modulator of food seeking. J Neurosci *24*, 1265-1271.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci *23*, 9913-9923.

Saunders, B.T., and Robinson, T.E. (2012). The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses. Eur J Neurosci *36*, 2521-2532.

Schmidt, R., Leventhal, D.K., Mallet, N., Chen, F., and Berke, J.D. (2013). Canceling actions involves a race between basal ganglia pathways. Nat Neurosci *16*, 1118-1124.

Schultz, W. (1997). A Neural Substrate of Prediction and Reward. Science *275*, 1593-1599.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J Neurophysiol *80*, 1-27.

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. Nature Reviews Neuroscience

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci *13*, 900-913.

Sulzer, D., Cragg, S.J., and Rice, M.E. (2016). Striatal dopamine neurotransmission: Regulation of release and uptake. Basal Ganglia *6*, 123-148.

Svoboda, K., Denk, W., Kleinfeld, D., and Tank, D.W. (1997). In vivo dendritic calcium dynamics in neocortical pyramidal neurons. Nature *385*, 161-165.

Tantama, M., Hung, Y.P., and Yellen, G. (2012). Optogenetic reporters: Fluorescent protein-based genetically encoded indicators of signaling and metabolism in the brain. Prog Brain Res *196*, 235-263.

Threlfell, S., and Cragg, S.J. (2011). Dopamine Signaling in Dorsal Versus Ventral Striatum: The Dynamic Role of Cholinergic Interneurons. Front Syst Neurosci *5*

Threlfell, S., Lalic, T., Platt, N.J., Jennings, K.A., Deisseroth, K., and Cragg, S.J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron *75*, 58-64.

Tian, L., Hires, S.A., Mao, T., Huber, D., Chiappe, M.E., Chalasani, S.H., Petreanu, L., Akerboom, J., McKinney, S.A., Schreiter, E.R., Bargmann, C.I., Jayaraman, V., Svoboda, K., and Looger, L.L. (2009). Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. Nat Methods *6*, 875-881.

Tobler, P.N., Fiorillo, C.D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. Science *307*, 1642-1645.

Tsien, R.W. (1983). Calcium channels in excitable cell membranes. Annual review of physiology *45*, 341-358.

Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A., and Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. Neuron *74*, 858-873.

Zalocusky, K.A., Ramakrishnan, C., Lerner, T.N., Davidson, T.J., Knutson, B., and Deisseroth, K. (2016). Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making. Nature *531*, 642-646.

CHAPTER 5: GENERAL DISCUSSION

Brain mechanisms of reinforcement learning and adaptive decision-making are widely accepted to critically involve the basal ganglia (BG) and the neurotransmitter dopamine (DA). Midbrain DA firing can encode errors in reward prediction, providing a learning signal to guide future behaviors. Yet, dopamine is also a key modulator of motivation, invigorating current behavior. Prevailing theories of DA emphasize its role in either affecting current performance, or modulating reward-related learning. This thesis is aimed at resolving exactly how dopamine makes *simultaneous* contributions to experimentally dissociable learning and motivational processes.

In chapter two, I described results from several experiments using a range of techniques directed at assessing the apparent simultaneous (and complementary) contributions of mesolimbic DA to learning and motivation. We reported that, across multiple timescales, DA release encodes an estimate of temporally discounted future rewards (a value function). This rapidly evolving decision variable is then employed for both learning and motivational functions.

Using microdialysis, we found that DA fluctuations on the order of minutes correlated with reward-rate. Sub-second voltammetric measurements of DA revealed that early-trial fluctuations reflect reward expectation, used for invigorating approach behaviors. Indeed, brief optogenetic stimulation of the VTA immediately motivated rats to perform the task, confirming that early-trial DA changes are causally involved in the decision to exert effort. During the feedback epoch (moments after outcome is revealed), abrupt deflections in striatal DA concentration correlated with degree of reward surprise, signaling reward prediction errors (RPE). Brief optogenetic stimulation

(or inhibition) of VTA neurons confirmed the causal role of these rapid DA deflections in bidirectional reinforcement of action-values to affect left/right choices on future trials.

Taken together, these findings elaborate on DA mechanisms for how rewards produce related but separable psychological processes of motivation and learning. Higher reward rates are associated with increases in overall ‘tonic’ striatal DA, and these increases have an activational function, for example, by elevating the hazard rate of task-engagement. DA changes that affect motivation, however, do not have to be inherently slow (like previously suggested (Niv, 2007; Niv et al., 2007)) because rapid increases in DA are sufficient to promote flexible approach. We suggested that DA’s influence over excitability of MSNs (Kitai and Surmeier, 1992; Surmeier et al., 1995) likely underlies action energization and invigoration. Some experimental evidence supports this notion; for example, MSN activity within the NAcc are associated with flexible approach behaviors (McGinty et al., 2013; Cheer et al., 2007; Day et al., 2006), and optogenetic stimulation of striatal dMSNs cause immediate locomotion (Kravitz et al., 2010; Tecuapetla et al., 2016; Yttri and Dudman, 2016; Roseberry et al., 2016). Furthermore, during the feedback epoch, surprising outcomes produced errors in reward prediction, which are coded by rapid deflections in DA value signal. These abrupt changes in DA affect D1/D2 receptor occupancy within the striatum (Richfield et al., 1989), and promote updating of cached state and action values via DA dependent cortico-striatal LTP (Maia and Frank, 2011; Shen et al., 2008; Yagishita et al., 2014). Indeed, many experiments in brain slices have characterized the cellular mechanisms of DA dependent plasticity (see Kreitzer and Malenka, 2008; Gerfen and Surmeier, 2011 for review), while others have reported in vivo cortico-striatal plasticity (Charpier and Deniau, 1997; Charpier et al., 1999; Reynolds and Wickens, 2000; Stoetzner et al., 2010) that compliment changes in learned behavior (Reynolds et al., 2001; Xiong et al., 2015).

Whether and how ‘phasic’ DA signals for prediction errors accumulate to influence ‘tonic' dopamine remains an understudied aspect of DA neurophysiology. The combination of baseline spiking and DA reuptake transporter (DAT) function are argued

to regulate the 'tonic' (ambient) levels of forebrain DA via release-reuptake equilibrium (Floresco et al., 2003; Shou et al., 2006). On the other hand, electrical activation of DA cells has the potential to drive efflux of >2uM striatal DA levels (Gonon et al., 1980; Gonon, 1997; Garris et al., 1999), but, physiologically (and behaviorally) relevant activation of DA cells (~30Hz bursts for ~300ms) likely produce ~50 nM of striatal DA transients (Aragona et al., 2008; Aragona et al., 2009; Wightman et al., 2007) that are rapidly cleared.

The prevailing quantification methods for forebrain DA levels during behavioral performance are in vivo microdialysis (Tidey and Miczek, 1996; Hernandez and Hoebel, 1988; Shou et al., 2006) and voltammetry (Garris et al., 1999; Phillips et al., 2003; Heien et al., 2005; Roitman et al., 2008; Clark et al., 2010). While both afford clear advantages, they are each imperfect. Microdialysis probes are large and sampling rates slow (usually on the order of a few minutes), making it best suited for the measurement of uniformly spread and slow varying DA. On the other had, voltammetry can detect sub-second changes in DA (localized to micro-domains surrounding 7x100um carbon fiber cylinder), but unstable probes make it susceptible to measurement drifts. So, voltammetry is a differential technique, and not ideally suited for comparing DA levels that span several minutes. The respective weaknesses of each technique make it difficult to experimentally test whether 'tonic' and 'phasic' DA release patterns are separate, computationally (and behaviorally) relevant 'channels' of DA release.

Naturally, DA release fluctuations are inherently fast (i.e. every exocytotic vesicular fusion event lasts a few milliseconds), but the temporal resolution of DA measurement has historically dictated the experimental interpretation underlying DA dynamics. For example, microdialysis results are reflexively interpreted as quantification of the 'tonic' levels whereas; voltammetric changes are referred as 'phasic' levels. This method-specific nomenclature has mostly set the basis for the notion that tonic and phasic changes in DA are separate signaling mechanisms of mesolimbic DA (Goto et al., 2007; Cragg and Rice, 2004; Phillips and Wightman, 2004; Floresco et al., 2003;

Wightman and Robinson, 2002; Niv, 2007; Niv et al., 2007; Dreyer et al., 2010). But, this notion remains to be clearly demonstrated experimentally.

Indeed, there are at least two clues that suggest ‘tonic’ and ‘phasic’ release are distinct processes. First, midbrain DA cells appear to fire in two modes, an irregular slow (tonic) rate and synchronous (phasic) bursts of action potentials (Grace and Bunney, 1984). These two firing modes are regulated by separate afferent inputs to DA cells (Floresco et al., 2003), and changes in the tonic firing rate(or more specifically, changes in the number of cells participating in irregular discharge) affect the ‘tonic’ levels of DA in the striatum measured via microdialysis. On the other hand, changes in the bursting of DA cells promote fast release of DA that are rapidly cleared from the cleft by DAT (Gonon, 1997){ (Garris et al., 1994; Cragg and Rice, 2004; Rice et al., 2011; Sulzer et al., 2016), and not observed to change ‘tonic’ levels via microdialysis (Floresco et al., 2003). Critically, increased phasic bursts are only associated with elevated ‘tonic’ DA only if DAT function is compromised, suggesting that DA spillover (i.e extra-synaptic DA) is the major contributor to ‘tonic’ levels. Indeed the consensus definition of ‘tonic’ DA is based on its capacity to escape from the cleft (and activate extra-synaptic receptors), and not dependent on the time-course of its variability or action (Gonon, 1997; Cragg and Rice, 2004; Sulzer and Pothos, 2000; Garris et al., 1994; Rice et al., 2011; Sulzer et al., 2016).

The second evidence for tonic/phasic distinction is the voltammetric detection of decreases in DA following optogenetic inhibition (McCutcheon et al., 2014) or due to physiological pauses in DA cell firing (McCutcheon et al., 2012; Roitman et al., 2008; Hamid et al., 2016). Rapid decrease in striatal DA from baseline levels following suppression of midbrain activity suggests that the tonic firing of DA cells contribute to some ambient forebrain level.

Some unexpected findings of our study are discussed in chapter 2. Notably, activation of DA cells early in trials immediately promotes approach behaviors, but does not appear to influence future selection of left/right choices. In addition, stimulation during the outcome epoch strengthens left/right choice associations, but

does not affect future motivated approach behaviors. While it is apparent that DA makes key contributions to motivated performance and learning action values, these observations provide clues for the temporal organization of DA mediated plasticity. I followed up these experiments in chapter 3.

Rat performance of our trial-and-error task likely involves motivational decisions ('engage task?') in addition to an operant choice ('which action?', i.e. left/right). These two processes are apparent from adaptive latencies to start task-performance (Figure 2.1g) and also from flexible choice behaviors (Figure 2.1d) that are sensitive to changes in reward probabilities for left/right actions. In chapter 3, I describe the time course of DA's role in reinforcing motivational and choice specific values. We found that optogenetic DA stimulation during the epochs of selecting/executing the left/right choice strengthen its associative value, making it more likely to be repeated. This temporal specificity of DA is consistent with the credit assignment notion in RL. Specifically, performed actions have a decaying eligibility trace for update by learning signals (Sutton and Barto, 2012; Pan et al., 2005; Redgrave et al., 2008). Thus, actions closely followed by rewards undergo the strongest reinforcement (Thorndike, 1911; Donahoe et al., 1993; Balleine and Dickinson, 1998; Black et al., 1985). Our finding suggests a neural mechanism of DA mediated 'plasticity windows' that potentiate temporally organized striatal MSN activity (Ito and Doya, 2009; Atallah et al., 2014; Ito and Doya, 2015).

Our optogenetic experiments from chapter 3 also suggest that DA *alone* are not sufficient to increase state-values. Specifically, Side-In stimulations (that reinforce choice behavior) did not affect latency on future trials (Figure 2.14b, bottom). This suggested that DA may reflect some, but not all reinforcing qualities of rewards, and may function together with other neuromodulatory systems to mediate learning. Within the striatum, DA is closely associated with local cholinergic signaling, and some have proposed that tonically active striatal cholinergic interneurons (TANs) regulate plasticity windows for DA mediated learning (Cragg, 2006; Threlfell and Cragg, 2011; Cohen and Frank, 2009; Franklin and Frank, 2015).

Acetylcholine(ACh) within the striatum is exclusively released by TANs, which, similar to DA axons, make extensive synaptic contacts. It is estimated that DA and TAN terminals are usually ~1μm apart (Descarries et al., 1996), setting a platform for crosstalk between the two neuromodulatory systems. TANs respond to rewards with a biphasic pause-excitation in firing (Ravel et al., 2001). These pauses during outcome epochs are coincident with DA-RPE signals (Morris et al., 2004), and some have suggested that extracellular ACh may set the necessary microcircuit dynamics for enhanced learning selectively during the outcome epoch (Cragg, 2006; Threlfell and Cragg, 2011; Cohen and Frank, 2009; Franklin and Frank, 2015).

In addition to regulating DA's influence over postsynaptic cells that underlie behavioral learning, ACh is one of many local modulators of terminal DA release, potentially decoupling exocytosis from cell body spiking. For instance, several investigators have argued that DA is a broadcast signal that relays a uniform message to downstream targets (Schultz, 1998; Glimcher, 2011; Kim et al., 2012; Eshel et al., 2016). But, experimentally-evoked and naturally occurring DA fluctuations are variable across the striatal micro (and macro) domains (Wightman et al., 2007; Aragona et al., 2009; Badrinarayan et al., 2012; Brown et al., 2011; Zhang et al., 2009). We now know that distinct subpopulations of DA cells i) receive different inputs and make diverse projections (Ikemoto, 2007; Watabe-Uchida et al., 2012; Lammel et al., 2008; Lerner et al., 2015; Wall et al., 2013), ii) vary in expression of ion channels (Lammel et al., 2008), iii) form diverse groups based on neurotransmitter co-release (Tritsch et al., 2012; Stuber et al., 2010; Zhang et al., 2015; Yamaguchi et al., 2007) and iv) are activated differently in response to reward or aversive cues (Cohen et al., 2012; Matsumoto and Hikosaka, 2009). Together these studies suggest that midbrain-forebrain DA axis is associated with inherent mechanisms for divergence of cell body spiking and terminal release patterns.

In chapter 4, I set out to assess: i) if VTA activity patterns are different from striatal release dynamics, ii) if such a divergence of signaling exists, what behaviorally relevant signals are transmitted at the respective node? iii) and finally, if TANs are

causally involved in meditating this divergent DA signaling on midbrain-forebrain axis. Some of my initial goals were not achieved partly due to technical difficulties and variability in collected data that preclude clear conclusions. I performed fiber photometry of midbrain DA cells to assay bulk activity of VTA during rat performance of three behavioral tasks. We found that VTA activity correlated with RPE signals, notably different from previous reports of value-coding in NAcc DA fluctuations. This indicates that midbrain activation and forebrain release patterns are not identical, and are potentially disjoint.

One likely mechanism for how DA release within the striatum is different from midbrain spiking involves cholinergic modulation of DA release within the striatum. Indeed, many studies have characterized the relationship of DA release and TAN activity in brain slice preparations (Threlfell et al., 2012; Straub et al., 2014; Wieland et al., 2014; Cachope et al., 2012; Nelson et al., 2014; Chuhma et al., 2014). As a result, there is a wealth of data, but a clear framework that describes DA-ACh interactions with a succinct causal directionality is thus far lacking. Nonetheless, the following three observations are very clear and replicated in different labs. 1) selective activation of TANs can directly cause synaptic efflux of DA from terminals, in the absence of cell body spiking (Threlfell et al., 2012; Cachope et al., 2012). 2) If DA terminals are activated at low frequencies (1-10 Hz), ACh augments DA release (Rice and Cragg, 2004; Threlfell et al., 2012; Cachope et al., 2012). 3) On the other hand, if DA terminals are activated at high frequencies (20-100 Hz), ACh release blunts DA release (Rice and Cragg, 2004; Threlfell et al., 2012; Cachope et al., 2012). Together these results suggest that TAN activation serves as a low pass filter for DA release, such that AChR receptor occupancy selectively amplifies DA release in response to low-frequency firing mode of DA cells. By contrast, brief TAN pauses enhance [DA] released in response to high frequency activation of DA cells while selectively attenuating DA release in response to low frequency firing. Put together, these studies indicate a cellular mechanism for the transformation of an cell body spiking RPE signal into modular release pattens, that can for example take the form of a motivationally relevant value function.

Some unresolved questions in this thesis could be addressed in future investigations. Our voltammetric observations were restricted to ventral striatum, and the dynamics of DA transmission, together with the decision variable signaled in the dorsal striatum must be assessed. Furthermore, temporally precise activation or suppression of DA transmission in striatal subregions could she light on the target-region specific functions of DA. Another important direction of future experimentation is disentangling the computational role of DA-ACh interactions. For example, one testable working hypothesis is that DA increases to reward-feedback cues require a coincident pause in accumbens TANs to mediate state-value updating. To test this, cholinergic interneurons would be transiently stimulated (or suppressed) randomly thought the task, as described in experiment-two of chapter 3. One prediction would be an increase in future trial latency only if the TAN pauses are occluded by ChR2 activation. In addition to testing the behavioral consequence of manipulating cholinergic interneurons, future studies could also characterize the shaping of DA signals by acetylcholine in vivo. Together with robust behavioral tasks and temporally precise manipulations, these experiments could identify DA dependent (transformation of signals) or independent (modulation of postsynaptic cell of the striatum) mechanisms of TAN influence over behavioral control.

In conclusion, I assessed the neuroeconomic signals relayed by forebrain DA levels, and exactly how it simultaneously mediates motivational arousal and learning. Mesolimbic DA is a critical component of brain circuits for action selection and learning from evaluative feedback. I have demonstrated that accumbens DA fluctuations encode the sum of discounted future rewards, a variable used for effort allocation and deciding to begin work. Rapid deviations from this moment-by-moment expectation signals reward prediction errors that serve to strengthen associations. Artificial increases in DA (even in the absence of rewards) are necessary and sufficient to strengthen the value of executed actions, suggesting that DA mediates key reinforcing qualities of natural rewards. On the other hand, other valuation systems (especially those that regulate coarse effort allocation and motivational arousal, i.e. state values) are not reinforced by artificial DA stimulations, and appear to require the recruitment of additional

components in the learning circuit. Lastly, the DA circuit architecture, including projection anatomy and integration into target microcircuit, is organized to efficiently control motivated behaviors. Specifically, some results including our own suggest that DA cell body spiking is not identical to the release patterns in forebrain target regions. I suggest that such a disparity in spiking and DA release is an adaptive feature for efficient reward-related computation, by allowing target regions to transform DA release into signals for region-specific control of goal-directed behaviors.

References

Aragona, B.J., Cleaveland, N.A., Stuber, G.D., Day, J.J., Carelli, R.M., and Wightman, R.M. (2008). Preferential enhancement of dopamine transmission within the nucleus accumbens shell by cocaine is attributable to a direct increase in phasic dopamine release events. J Neurosci *28*, 8821-8831.

Aragona, B.J., Day, J.J., Roitman, M.F., Cleaveland, N.A., Wightman, R.M., and Carelli, R.M. (2009). Regional specificity in the real-time development of phasic dopamine transmission patterns during acquisition of a cue-cocaine association in rats. Eur J Neurosci *30*, 1889-1899.

Atallah, H., McCool, A., Howe, M., and Graybiel, A. (2014). Neurons in the Ventral Striatum Exhibit Cell-Type-Specific Representations of Outcome during Learning. Neuron *82*, 1145-1156.

Badrinarayan, A., Wescott, S.A., Weele, C.M.V., Saunders, B.T., Couturier, B.E., Maren, S., and Aragona, B.J. (2012). Aversive Stimuli Differentially Modulate Real-Time Dopamine Transmission Dynamics within the Nucleus Accumbens Core and Shell. J Neurosci *32*, 15779-15790.

Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407-419.

Black, J., Belluzzi, J.D., and Stein, L. (1985). Reinforcement delay of one second severely impairs acquisition of brain self-stimulation. Brain Res *359*, 113-119.

Brown, H.D., McCutcheon, J.E., Cone, J.J., Ragozzino, M.E., and Roitman, M.F. (2011). Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. Eur J Neurosci *34*, 1997-2006.

Cachope, R., Mateo, Y., Mathur, B.N., Irving, J., Wang, H.-L., Morales, M., Lovinger, D.M., and Cheer, J.F. (2012). Selective activation of cholinergic interneurons enhances

accumbal phasic dopamine release: setting the tone for reward processing. Cell Rep *2*, 33-41.

Charpier, S., and Deniau, J.M. (1997). In vivo activity-dependent plasticity at cortico-striatal connections: evidence for physiological long-term potentiation. Proceedings of the National Academy of Sciences *94*, 7036-7040.

Charpier, S., Mahon, S., and Deniau, J.-M. (1999). In vivo induction of striatal long-term potentiation by low-frequency stimulation of the cerebral cortex. Neuroscience *91*, 1209-1222.

Cheer, J.F., Aragona, B.J., Heien, M.L., Seipel, A.T., Carelli, R.M., and Wightman, R.M. (2007). Coordinated accumbal dopamine release and neural activity drive goal-directed behavior. Neuron *54*, 237-244.

Chuhma, N., Mingote, S., Moore, H., and Rayport, S. (2014). Dopamine neurons control striatal cholinergic neurons via regionally heterogeneous dopamine and glutamate signaling. Neuron *81*, 901-912.

Clark, J.J., Sandberg, S.G., Wanat, M.J., Gan, J.O., Horne, E.A., Hart, A.S., Akers, C.A., Parker, J.G., Willuhn, I., and Martinez, V. (2010). Chronic microsensors for longitudinal, subsecond dopamine detection in behaving animals. Nat Methods *7*, 126-129.

Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature *482*, 85-88.

Cohen, M.X., and Frank, M.J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. Behav Brain Res *199*, 141-156.

Cragg, S.J. (2006). Meaningful silences: how dopamine listens to the ACh pause. Trends Neurosci *29*, 125-131.

Cragg, S.J., and Rice, M.E. (2004). DAncing past the DAT at a DA synapse. Trends Neurosci *27*, 270-277.

Day, J.J., Wheeler, R.A., Roitman, M.F., and Carelli, R.M. (2006). Nucleus accumbens neurons encode Pavlovian approach behaviors: evidence from an autoshaping paradigm. Eur J Neurosci *23*, 1341-1351.

Descarries, L., Watkins, K.C., Garcia, S., Bosler, O., and Doucet, G. (1996). Dual character, asynaptic and synaptic, of the dopamine innervation in adult rat neostriatum: a quantitative autoradiographic and immunocytochemical analysis. Journal of Comparative Neurology *375*, 167-186.

Donahoe, J.W., Burgos, J.E., and Palmer, D.C. (1993). A selectionist approach to reinforcement. Journal of the experimental analysis of behavior *60*, 17-40.

Dreyer, J.K., Herrik, K.F., Berg, R.W., and Hounsgaard, J.D. (2010). Influence of phasic and tonic dopamine release on receptor activation. The Journal of Neuroscience *30*, 14273-14283.

Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. Nat Neurosci *19*, 479-486.

Floresco, S.B., West, A.R., Ash, B., Moore, H., and Grace, A.A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. Nat Neurosci *6*, 968-973.

Franklin, N.T., and Frank, M.J. (2015). A cholinergic feedback circuitto regulate striatal population uncertainty and optimize reinforcement learning. eLife , e12029.

Garris, P.A., Ciolkowski, E.L., Pastore, P., and Wightman, R.M. (1994). Efflux of dopamine from the synaptic cleft in the nucleus accumbens of the rat brain. The Journal of neuroscience *14*, 6084-6093.

Garris, P.A., Kilpatrick, M., Bunin, M.A., Michael, D., Walker, Q.D., and Wightman, R.M. (1999). Dissociation of dopamine release in the nucleus accumbens from intracranial self-stimulation. Nature *398*, 67-69.

Gerfen, C.R., and Surmeier, D.J. (2011). Modulation of striatal projection systems by dopamine. Annu Rev Neurosci *34*, 441-466.

Glimcher, P.W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A *108 Suppl 3*, 15647-15654.

Gonon, F. (1997). Prolonged and extrasynaptic excitatory action of dopamine mediated by D1 receptors in the rat striatum in vivo. The Journal of neuroscience *17*, 5972-5978.

Gonon, F., Buda, M., Cespuglio, R., Jouvet, M., and Pujol, J.F. (1980). In vivo electrochemical detection of catechols in the neostriatum of anaesthetized rats: dopamine or DOPAC? Nature *286*, 902-904.

Goto, Y., Otani, S., and Grace, A.A. (2007). The Yin and Yang of dopamine release: a new perspective. Neuropharmacology *53*, 583-587.

Grace, A.A., and Bunney, B.S. (1984). The control of firing pattern in nigral dopamine neurons: single spike firing. The Journal of neuroscience *4*, 2866-2876.

Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., and Berke, J.D. (2016). Mesolimbic dopamine signals the value of work. Nat Neurosci *19*, 117-126.

Heien, M.L., Khan, A.S., Ariansen, J.L., Cheer, J.F., Phillips, P.E.M., Wassum, K.M., and Wightman, R.M. (2005). Real-time measurement of dopamine fluctuations after cocaine in the brain of behaving rats. Proc Natl Acad Sci U S A *102*, 10023-10028.

Hernandez, L., and Hoebel, B.G. (1988). Food reward and cocaine increase extracellular dopamine in the nucleus accumbens as measured by microdialysis. Life sciences *42*, 1705-1712.

Ikemoto, S. (2007). Dopamine reward circuitry: two projection systems from the ventral midbrain to the nucleus accumbens-olfactory tubercle complex. Brain Res Rev *56*, 27-78.

Ito, M., and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci *29*, 9861-9874.

Ito, M., and Doya, K. (2015). Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. J Neurosci *35*, 3499-3514.

Kim, Y., Wood, J., and Moghaddam, B. (2012). Coordinated activity of ventral tegmental neurons adapts to appetitive and aversive learning. PLoS One *7*, e29766.

Kitai, S.T., and Surmeier, D.J. (1992). Cholinergic and dopaminergic modulation of potassium conductances in neostriatal neurons. Advances in neurology *60*, 40-52.

Kravitz, A.V., Freeze, B.S., Parker, P.R., Kay, K., Thwin, M.T., Deisseroth, K., and Kreitzer, A.C. (2010). Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. Nature *466*, 622-626.

Kreitzer, A.C., and Malenka, R.C. (2008). Striatal plasticity and basal ganglia circuit function. Neuron *60*, 543-554.

Lammel, S., Hetzel, A., Häckel, O., Jones, I., Liss, B., and Roeper, J. (2008). Unique properties of mesoprefrontal neurons within a dual mesocorticolimbic dopamine system. Neuron *57*, 760-773.

Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K., Malenka, R.C., Luo, L., Tomer, R., and Deisseroth, K. (2015). Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell *162*, 635-647.

Maia, T.V., and Frank, M.J. (2011). From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci *14*, 154-162.

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837-841.

McCutcheon, J.E., Cone, J.J., Sinon, C.G., Fortin, S.M., Kantak, P.A., Witten, I.B., Deisseroth, K., Stuber, G.D., and Roitman, M.F. (2014). Optical suppression of drug-evoked phasic dopamine release. Front Neural Circuits *8*

McCutcheon, J.E., Ebner, S.R., Loriaux, A.L., and Roitman, M.F. (2012). Encoding of aversion by dopamine and the nucleus accumbens. Front Neurosci *6*, 137.

McGinty, V.B., Lardeux, S., Taha, S.A., Kim, J.J., and Nicola, S.M. (2013). Invigoration of reward seeking by cue and proximity encoding in the nucleus accumbens. Neuron *78*, 910-922.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., and Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron *43*, 133-143.

Niv, Y. (2007). Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation? Ann N Y Acad Sci *1104*, 357-376.

Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl) *191*, 507-520.

Onge, J.R.S., Ahn, S., Phillips, A.G., and Floresco, S.B. (2012). Dynamic fluctuations in dopamine efflux in the prefrontal cortex and nucleus accumbens during risk-based decision making. The Journal of Neuroscience *32*, 16880-16891.

Pan, W.X., Schmidt, R., Wickens, J.R., and Hyland, B.I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J Neurosci *25*, 6235-6242.

Phillips, P.E., and Wightman, R.M. (2004). Extrasynaptic dopamine and phasic neuronal activity. Nat Neurosci *7*, 199-199.

Phillips, P.E., Stuber, G.D., Heien, M.L., Wightman, R.M., and Carelli, R.M. (2003). Subsecond dopamine release promotes cocaine seeking. Nature *422*, 614-618.

Ravel, S., Sardo, P., Legallet, E., and Apicella, P. (2001). Reward unpredictability inside and outside of a task context as a determinant of the responses of tonically active neurons in the monkey striatum. The Journal of neuroscience *21*, 5730-5739.

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? Brain Res Rev *58*, 322-339.

Reynolds, J.N., Hyland, B.I., and Wickens, J.R. (2001). A cellular mechanism of reward-related learning. Nature *413*, 67-70.

Reynolds, J.N.J., and Wickens, J.R. (2000). Substantia nigra dopamine regulates synaptic plasticity and membrane potential fluctuations in the rat neostriatum, in vivo. Neuroscience *99*, 199-203.

Rice, M.E., and Cragg, S.J. (2004). Nicotine amplifies reward-related dopamine signals in striatum. Nat Neurosci *7*, 583-584.

Rice, M.E., Patel, J.C., and Cragg, S.J. (2011). Dopamine release in the basal ganglia. Neuroscience *198*, 112-137.

Richfield, E.K., Penney, J.B., and Young, A.B. (1989). Anatomical and affinity state comparisons between dopamine D 1 and D 2 receptors in the rat central nervous system. Neuroscience *30*, 767-777.

Roitman, M.F., Wheeler, R.A., Wightman, R.M., and Carelli, R.M. (2008). Real-time chemical responses in the nucleus accumbens differentiate rewarding and aversive stimuli. Nat Neurosci *11*, 1376-1377.

Roseberry, T.K., Lee, A.M., Lalive, A.L., Wilbrecht, L., Bonci, A., and Kreitzer, A.C. (2016). Cell-Type-Specific Control of Brainstem Locomotor Circuits by Basal Ganglia. Cell *164*, 526-537.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J Neurophysiol *80*, 1-27.

Shen, W., Flajolet, M., Greengard, P., and Surmeier, D.J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. Science *321*, 848-851.

Shou, M., Ferrario, C.R., Schultz, K.N., Robinson, T.E., and Kennedy, R.T. (2006). Monitoring dopamine in vivo by microdialysis sampling and on-line CE-laser-induced fluorescence. Analytical chemistry *78*, 6717-6725.

Stoetzner, C.R., Pettibone, J.R., and Berke, J.D. (2010). State-dependent plasticity of the corticostriatal pathway. Neuroscience *165*, 1013-1018.

Straub, C., Tritsch, N.X., Hagan, N.A., Gu, C., and Sabatini, B.L. (2014). Multiphasic modulation of cholinergic interneurons by nigrostriatal afferents. J Neurosci *34*, 8557-8569.

Stuber, G.D., Hnasko, T.S., Britt, J.P., Edwards, R.H., and Bonci, A. (2010). Dopaminergic terminals in the nucleus accumbens but not the dorsal striatum corelease glutamate. J Neurosci *30*, 8229-8233.

Sulzer, D., and Pothos, E.N. (2000). Regulation of quantal size by presynaptic mechanisms. Reviews in the neurosciences *11*, 159-212.

Sulzer, D., Cragg, S.J., and Rice, M.E. (2016). Striatal dopamine neurotransmission: Regulation of release and uptake. Basal Ganglia *6*, 123-148.

Surmeier, D.J., Bargas, J., Hemmings, H.C., Nairn, A.C., and Greengard, P. (1995). Modulation of calcium currents by a D1 dopaminergic protein kinase/phosphatase cascade in rat neostriatal neurons. Neuron *14*, 385-397.

Sutton, R.S., and Barto, A.G. (2012). Reinforcement learning: An introduction (Cambridge Univ Press).

Tecuapetla, F., Jin, X., Lima, S.Q., and Costa, R.M. (2016). Complementary Contributions of Striatal Projection Pathways to Action Initiation and Execution. Cell *166*, 703-715.

Thorndike, E.L. (1911). Animal intelligence: Experimental studies (Transaction Publishers).

Threlfell, S., and Cragg, S.J. (2011). Dopamine Signaling in Dorsal Versus Ventral Striatum: The Dynamic Role of Cholinergic Interneurons. Front Syst Neurosci *5*

Threlfell, S., Lalic, T., Platt, N.J., Jennings, K.A., Deisseroth, K., and Cragg, S.J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron *75*, 58-64.

Tidey, J.W., and Miczek, K.A. (1996). Social defeat stress selectively alters mesocorticolimbic dopamine release: an in vivo microdialysis study. Brain Res *721*, 140-149.

Tritsch, N.X., Ding, J.B., and Sabatini, B.L. (2012). Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. Nature *490*, 262-266.

Wall, N.R., De La Parra, M., Callaway, E.M., and Kreitzer, A.C. (2013). Differential innervation of direct- and indirect-pathway striatal projection neurons. Neuron *79*, 347-360.

Watabe-Uchida, M., Zhu, L., Ogawa, S.K., Vamanrao, A., and Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. Neuron *74*, 858-873.

Wieland, S., Du, D., Oswald, M.J., Parlato, R., Köhr, G., and Kelsch, W. (2014). Phasic Dopaminergic Activity Exerts Fast Control of Cholinergic Interneuron Firing via Sequential NMDA, D2, and D1 Receptor Activation. The Journal of Neuroscience *34*, 11549-11559.

Wightman, R.M., and Robinson, D.L. (2002). Transient changes in mesolimbic dopamine and their association with reward. Journal of neurochemistry *82*, 721-735.

Wightman, R.M., Heien, M.L., Wassum, K.M., Sombers, L.A., Aragona, B.J., Khan, A.S., Ariansen, J.L., Cheer, J.F., Phillips, P.E., and Carelli, R.M. (2007). Dopamine release is heterogeneous within microenvironments of the rat nucleus accumbens. Eur J Neurosci *26*, 2046-2054.

Xiong, Q., Znamenskiy, P., and Zador, A.M. (2015). Selective corticostriatal plasticity during acquisition of an auditory discrimination task. Nature

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science *345*, 1616-1620.

Yamaguchi, T., Sheen, W., and Morales, M. (2007). Glutamatergic neurons are present in the rat ventral tegmental area. European Journal of Neuroscience *25*, 106-118.

Yttri, E.A., and Dudman, J.T. (2016). Opponent and bidirectional control of movement velocity in the basal ganglia. Nature *533*, 402-406.

Zhang, L., Doyon, W.M., Clark, J.J., Phillips, P.E., and Dani, J.A. (2009). Controls of tonic and phasic dopamine transmission in the dorsal and ventral striatum. Mol Pharmacol *76*, 396-404.

Zhang, S., Qi, J., Li, X., Wang, H.-L., Britt, J.P., Hoffman, A.F., Bonci, A., Lupica, C.R., and Morales, M. (2015). Dopaminergic and glutamatergic microdomains in a subset of rodent mesoaccumbens axons. Nat Neurosci *18*, 386-392.