Mesolimbic dopamine transmission during decisions involving cost-benefit tradeoffs

Nicholas Garber Hollon

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Paul E. M. Phillips, Chair

Larry S. Zweifel

John F. Neumaier

Program Authorized to Offer Degree:

Neuroscience

University of Washington

Abstract

Mesolimbic dopamine transmission during decisions involving cost-benefit tradeoffs

Nick Garber Hollon

Chair of the Supervisory Committee:

Paul E. M. Phillips, Associate Professor

Department of Psychiatry and Behavioral Sciences

Real-world decisions frequently involve tradeoffs between multiple economic dimensions, and the integration of benefits and costs into a common currency of subjective value is fundamental to action selection. Phasic dopamine is widely regarded as a critical teaching signal for learning the values assigned to actions, and these stored ("cached") values can be read out from the dopamine response to the unexpected presentation of reward-predictive cues. In Chapter 1, I introduce current theories of the neural basis of economic decision making and focus on how phasic dopamine is thought to contribute to these processes.

In Chapter 2, I present data testing the critical question of whether dopamine-associated cached values align with animals' subjective preferences in a decision-making task involving cost-

benefit tradeoffs. Here, I observed a significant inversion between animals' behavioral preferences and the rank ordering of dopamine-reported cached values, indicating that these cached values cannot be the sole determinant of choices in simple economic decision making. These data challenge the fundamental tenet of contemporary theories of decision making which posit that dopamine-associated cached values are sufficient to serve as the basis for action selection.

In Chapter 3, I examine the question of which variant of reinforcement learning algorithms is instantiated by phasic dopamine transmission, as previous reports in the literature have arrived at conflicting conclusions. Using the cost-benefit tradeoff task described in Chapter 2, I analyzed what cue-evoked dopamine encodes when animals face choices between concurrently available options, finding that cue-evoked dopamine signals the cached value of the chosen option. Consistent with the notion that phasic dopamine transmission enacts an "on-policy" reinforcement learning algorithm that updates based on chosen state-action values, these results indicate that cue-evoked dopamine reflects post-decision information about the expected value of the outcome of an action that has already been selected via other neural substrates.

In Chapter 4, I conclude by summarizing and integrating these results, discussing how the contributions of phasic dopamine to cost-benefit decision making are more limited and perhaps more nuanced than previously thought. I extend this discussion to describe other neural substrates which might contribute to the aspects of decision making not accommodated by phasic dopamine transmission.

Table of Contents

List of Figures

List of Tables

Acknowledgments

There are many people I would like to thank for all their guidance and support throughout my time in graduate school: my parents, Steve Hollon and Judy Garber; my supervisor, Paul Phillips; my committee, Larry Zweifel, John Neumaier, Jeansok Kim, Michael Shadlen, and Sheri Mizumori; Mark Walton and Jeremy Clark for additional mentorship; Jerylin Gan and Monica Arnold, my collaborators throughout these experiments; our undergraduate assistants, Kaija Reinelt, Kevin Dofredo, and Madison Hatfield; Scott Ng-Evans for all his invaluable technical support; Christina Akers Sanford and all the work-study students in the PEMPlab who followed in keeping the lab running; all the other troublemakers in the PEMPlab over the years, Matthew Wanat, Stefan Sandberg, Ingo Willuhn, Vicente Martinez, Andrew Hart, Julia Lemos, and Lauren Burgeno; Geoffrey Boynton, Greg Horwitz, and Rajesh Rao for advice with data analysis and theoretical interpretation; Ann Wilkinson and Lucia Wisdom in the NeuBeh office, and the program directors, Jane Sullivan and David Perkel; Andrea Duran, to whom this dissertation is dedicated, and all our friends for helping me stay sane by wandering around outside with me, chasing pieces of plastic I threw, watching countless hours of football, hosting holiday feasts, and various other adventures.

Dedication

This dissertation is dedicated to Andrea Duran, my ladyhalf, for hanging in there with me all these years despite my constantly bringing my work home with me.

Chapter 1

Introduction

"We now have all the moving parts." – Paul Glimcher (2009).

The "Godfather of Neuroeconomics" concluded his presentation at the "Basal Ganglia in Health and Disease" symposium at MIT's McGovern Institute in 2009 by declaring that researchers had identified all the neural substrates and computational processes required to explain how organisms learn the subjective value of their actions and make choices based upon these learned values. As 2009 also was the year I began graduate school, this seemed a most opportune time to begin capitalizing on everything we now so fully understood about the neurobiology of value-guided decision making. There were numerous comprehensive reviews (Sugrue et al., 2005; Daw and Doya, 2006; Rangel et al., 2008; Cohen and Frank, 2009; Kable and Glimcher, 2009; van der Meer and Redish, 2010; Lee et al., 2012) and even a neuroeconomics textbook (Glimcher et al., 2009) replete with similar models and schematics detailing the neural basis of value-guided decision making. Indeed, many prominent researchers in this field began to set their sights on the aberrant decision making often observed in various neuropsychiatric disorders, with the goals of applying these neuroeconomic theories and computational models to gain better understanding of the disorders' etiology, develop more precise clinical assessment tools, and ultimately inform better treatments for these disorders (Redish, 2004; Kishida et al., 2010; Maia and Frank, 2011; Montague et al., 2012; Huys et al., 2013; Lee, 2013). As I transitioned into my graduate work in the Phillips lab, however, it was becoming increasingly apparent that the prevailing account in the field was less complete than was being advocated.

In this introductory chapter of my dissertation, I summarize this dominant model of value-guided decision making and its underlying neurobiology (section 1.1), highlighting the prevailing

hypothesis that rapid signaling by the neuromodulator dopamine plays a critical role in learning the subjective value of actions such that these dopamine-associated "cached" values determine animals' subsequent choices. I then review the initial evidence that led me to question this general premise (section 1.2) and identify an unresolved discrepancy in the literature regarding the specific variant of reinforcement learning algorithms that dopamine transmission seems to instantiate (section 1.3). In Chapter 2, I present results that challenge a fundamental premise of these current theories of value-guided decision making: namely, the general principle that these dopamine-associated cached values are sufficient as the basis for economic decisions (Hollon et al., 2014). In Chapter 3, I present data that address the unresolved discrepancy detailed in section 1.3. In the closing chapter, I discuss the implications of these findings for updating our current understanding of the neural basis of decision making and for identifying which "moving parts" we are still missing.

1.1) Contemporary theories of the neural basis of value-guided decision making

Our current state of knowledge regarding the neurobiology of value-guided decision making has been built upon a foundation of work in economics, psychology, machine learning, and neuroscience, with each contributing a rich tradition of theory and methodology that continues to influence ongoing research. The following sections briefly highlight the high-level economic and psychological framework for valuation and choice (*1.1a*), provide an overview of computational algorithms developed in machine learning (*1.1b*), and summarize recent evidence for how these may be implemented at the neurobiological level (*1.1c*).

1.1a) Economic and psychological framework

Work in economics spanning several centuries continues to inform contemporary neuroeconomic approaches to the investigation of decision making (Caplin and Glimcher, 2014). Much of this economic tradition has culminated in the axiomatic formalisms of neoclassical economics (Samuelson, 1937; von Neumann and Morgenstern, 1944), though modern neuroeconomics also incorporates more recent empirical observations from behavioral economics that describe limits to the normative neoclassical theories (Kahneman and Tversky, 1979; Tversky and Kahneman, 1981). Briefly, rational choice theory is built upon the assumption that individuals behave so as to maximize benefits and minimize costs. The central construct of "utility" is used to describe the overall desirability of some good or the satisfaction obtained from some action. Although this intuitive definition captures the subjective aspects of valuation, utility functions are formally assessed based on the observable choices of individuals in accord with the notion of "revealed preferences" – directly measuring choices provides a metric of the ordinal ranking of options relative to each other. Whereas neoclassical economic theories are built upon this premise that agents act "as if" maximizing utility (Friedman, 1953), within these theories utility itself remains an abstract construct derived from the choices it is used to explain. Neuroeconomic theories make the more explicit claim that utility has a physical basis that is mechanistically involved in determining choices, in that organisms' brains contain actual representations of the subjective values of options and use these cardinal subjective value representations as the basis for

comparing and ultimately selecting an option (Kable and Glimcher, 2009; Padoa-Schioppa, 2011; Lee et al., 2012), as detailed in section *1.1c* below.

These modern theories of the neural basis of value-guided decision making also grew out of parallel developments in psychology throughout the past century. Whereas the economic models focus on how individuals make choices based upon existing subjective valuations, this psychology literature remains central to informing current accounts of how organisms learn these subjective values in the first place. The computational algorithms for learning subjective value, described below in section *1.1b*, originally were developed as quantitative models of classical or Pavlovian conditioning, through which initially neutral stimuli are associated with biologically relevant outcomes (Pavlov, 1927), and instrumental or operant conditioning, through which animals learn from experience to repeat actions yielding satisfactory outcomes and refrain from actions producing discomfort (Thorndike, 1911). By operationalizing the concepts of reinforcement and punishment as powerful controllers of behavior (Skinner, 1938), the behaviorist tradition explicitly eschewed introspective constructs and instead favored strictly observable behavior in a manner that ultimately aligns well with the revealed preferences approach noted above. Work in recent decades has moved beyond the radical behaviorist framework and has developed more sophisticated behavioral procedures for investigating animal learning and cognition (Tolman, 1949; Bolles, 1972; Bindra, 1974; Dickinson, 1985; Colwill and Rescorla, 1986; Mackintosh, 1994; Toates, 1998; Packard, 2009), which increasingly have come to be used in conjunction with similarly advancing techniques for probing and perturbing the neural substrates underlying these psychological processes.

Overall, contemporary theories of the neural basis of economic decision making attempt to explain how organisms learn the subjective value of their actions and make choices based upon these subjective values. Researchers have examined neural substrates underlying decisions involving a variety of economic attributes including reward magnitude and probability (the product of which comprises the objective expected value), risk preference, temporal discounting, effort, aversiveness, and other factors influencing individuals' preferences. While some early work within this nascent field of neuroeconomics proposed that separable valuation systems (derived from behavioral economic models) compete for control of decision making (McClure et al., 2004; 2007), these accounts initially fell out of favor with more vocal neuroeconomics researchers who instead advocated a more monolithic (though distributed) system for ultimately representing a unitary, integrated subjective value scale as a common currency for choice (Schultz, 2006; Kable and Glimcher, 2007, 2009; Padoa-Schioppa, 2011; Schultz, 2013; Cai and Padoa-Schioppa, 2014; Lak et al., 2014; Rangel and Clithero, 2014; Stauffer et al., 2014). By definition, as a common currency for choice, these subjective values are said to be "domain general" (Padoa-Schioppa, 2011), meaning that they incorporate any and all factors (e.g., expected value, risk, delay, effort costs, etc.) exactly to the extent that these factors influence organisms' behavioral preferences. This common-neural-currency account represents the dominant framework that provides the starting basis for my experiments, as detailed in the following sections on computational algorithms for learning subjective values (*1.1b*) and the purported neural instantiation of these algorithms (*1.1c*). Nevertheless, in recent years there also has been a growing appreciation of and emphasis on a different sort of multiple-systems approach, in this case one derived from the machine learning literature; indeed, many of the same computationally-oriented theorists responsible for the prominence of the learning

algorithms described below within neuroeconomics have more recently directed their efforts towards understanding the interplay between so-called "model-free" and "model-based" systems within the brain (Daw et al., 2005; Gläscher et al., 2010; Daw et al., 2011; Dolan and Dayan, 2013; Daw and O'Doherty, 2014; Lee et al., 2014). The sections below focus on the model-free computational algorithms that thus far have been most influential in the neuroeconomic literature and therefore have provided the more immediate rationale for my experiments; within my concluding chapter, I return to a discussion of "model-based" valuation systems to address my results within the context of multiple valuation systems as well.

1.1b) Computational algorithms

Like the economic and psychological foundations described above, the computational algorithms commonly employed to model value-guided decision making entail the learning of value functions for use in action selection, with an agent's defined goal being the maximization of long-term reward from these actions. Note that the semantics used in texts describing these algorithms typically employ generic terms such as "reward" and "value." Current neuroeconomic theories spanning both the economic and machine learning traditions have made the explicit link between the economic concept of utility and the value functions of these learning algorithms (Kable and Glimcher, 2009; Lee et al., 2012); each term is used within its respective field to denote a common currency providing a basis upon which options can be compared and chosen.

The temporal difference reinforcement learning (TDRL) algorithms are a family of models in which agents learn through experience which states or actions yield reward (Sutton and Barto, 1998), where a "state" refers to a given configuration of perceived stimuli in the agent's environment. The original formulation of temporal-difference learning described learning to predict the value of states $V(s)$ (Sutton, 1988), thereby elaborating on earlier models for learning the strength of Pavlovian associations between conditioned and unconditioned stimuli (Bush and Mosteller, 1951; Rescorla and Wagner, 1972). The most notable new feature of the temporal-difference model was its updating of value functions at each arbitrarily small time step rather than just once at the end of a "trial" – indeed, the concept of a trial is not a necessary component of TDRL algorithms, instead simply being replaced by the notion that the same state will be repeatedly experienced at various moments in time throughout learning.

Each unique state s has an associated value function $V(s)$, referred to as that state's "cached" value, representing the expected value E of all rewards r predicted from that state from the current time t forward:

$$V(s_t) = E\,[\, r(s_t) + \gamma r(s_{t+1}) + \gamma^2 r(s_{t+2}) + \gamma^3 r(s_{t+3}) + \ldots \mid s_t \,] \tag{1.1}$$

where $0 < \gamma \leq 1$ is a discount factor that results in the exponential discounting of predicted future rewards. Importantly, the cached value function at the next immediate time step $V(s_{t+1})$ can be expressed the same way:

$$V(s_{t+1}) = E\,[\, r(s_{t+1}) + \gamma r(s_{t+2}) + \gamma^2 r(s_{t+3}) + \gamma^3 r(s_{t+4}) + \ldots \mid s_{t+1} \,] \tag{1.2}$$

This results in the ability to simplify the summation within $V(s_t)$ based upon $V(s_{t+1})$ in the following manner:

$$V(s_t) = E\,[\, r(s_t) + \gamma V(s_{t+1}) \mid s_t \,] \tag{1.3}$$

Whenever the agent transitions into a given state, a "prediction error" is generated according to the difference between these two sides of equation 1.3:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{1.4}$$

By effectively signaling any change in expectation of future reward and any received reward that differed in value from that expected, this temporal-difference prediction error serves as the critical teaching term used to update the cached value for that state:

$$V(s_t)_{\text{new}} = V(s_t)_{\text{old}} + \eta\delta_t \tag{1.5}$$

where $0 < \eta \leq 1$ is the learning rate parameter specifying how much each experienced iteration affects the cached value function. This temporal-difference learning system is said to be "model-free" because the agent is simply representing and updating these cached state values at each time point based on its history of experienced reward values without learning any internal model of the environment's transition probabilities between sequential states.

Beyond the purely predictive state values, there also are several TDRL variants designed for learning to select actions that will lead to reward in a manner more akin to instrumental conditioning, including the actor-critic method, Q-learning, and SARSA (i.e., "state-action-reward-state-action") (Watkins, 1989; Rummery and Niranjan, 1994; Sutton and Barto, 1998). The actor-critic model is comprised of a "critic" component, which learns state values $V(s)$ using temporal-difference prediction errors as detailed above (Equations 1.4-1.5), and a separate "actor" component, which learns and implements an action selection policy $\pi(s,a)$ representing the probability distribution over selecting each available action from a given state. This policy also is iteratively learned through an update rule based on the same temporal-difference prediction errors:

$$\pi(s_t, a_t)_{\text{new}} = \pi(s_t, a_t)_{\text{old}} + \eta\delta_t \tag{1.6}$$

Rather than learning separate state values and action selection policies, both Q-learning and SARSA entail direct learning of the value of actions taken from a given state—the state-action value $Q(s,a)$. These cached state-action values are analogous to the state values (Equation 1.1) above:

$$Q(s_t, a_t) = E\,[\, r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma^2 r(s_{t+2}, a_{t+2}) + \gamma^3 r(s_{t+3}, a_{t+3}) + \ldots \mid s_t, a_t] \tag{1.7}$$

However, each algorithm uses a slightly different form of the update rule. Q-learning represents and updates based on the greatest cached value available in a given state regardless of the actual action taken, and is therefore referred to as an "off-policy" algorithm:

$$\delta_t = r_t + \max_a [\gamma Q(s_{t+1}, a)] - Q(s_t, a_t) \quad (1.8)$$

SARSA is instead an "on-policy" algorithm, as it represents and updates based on the cached value of the action actually selected from that state:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \quad (1.9)$$

The experiments described in Chapter 3 aim to differentiate between these TDRL algorithms using neurobiological readouts of temporal-difference-prediction-error signals as described in section *1.1c* below.

In either algorithm for learning state-action values, these cached values are used to select actions according to a specified action selection rule. These selection rules differ in how they balance the tradeoff between exploration versus exploitation. A "greedy" selection rule always selects the action with the greatest cached value, resulting in exploitation of the action with the highest estimated value but at the expense of never exploring other options that might actually lead to greater reward. A near-greedy, "ε-greedy," rule entails selecting the action with the greatest cached value most of the time, but sampling alternative actions instead with some smaller probability ε, where the alternative action is selected at random in a uniform manner that is independent of its cached value. A third selection rule is the "softmax" method, in which the probability of selecting each action is weighted by the magnitude of its cached value:

$$P(a) = e^{Q(s,a)/\beta} / \Sigma_{i=1:n} [e^{Q(s,ai)/\beta}] \quad (1.10)$$

The temperature parameter β affects the slope of the resulting sigmoid curve, effectively balancing between exploration and exploitation. Softmax frequently has been found to provide the best fit to animals' actual choices and therefore is the most widely used choice rule in the neuroeconomics literature when modeling action selection (Sugrue et al., 2005; Daw and Doya, 2006; Cohen and Frank, 2009; Kable and Glimcher, 2009; Lee et al., 2012); accordingly, a softmax selection rule will be assumed throughout this dissertation, although the conclusions from my experiments generalize beyond this particular selection rule. In all cases within these models, the cached values serve as the sole basis for choice, with the action with the greatest cached value being selected most (if not all) of the time. This general principle yields the straightforward prediction tested in the experiments of Chapter 2: if choices are based solely on the relative magnitude of the cached values of available options and if we can obtain a neural readout of these cached values, then the option measured to have the greatest cached value should be the one that animals select most frequently. The section below outlines current evidence supporting hypotheses about the neurobiological implementation of TDRL-like algorithms for learning and action selection.

1.1c) Neurobiological implementation

At the core of contemporary theories of value-guided decision making is the observation that many dopamine-containing neurons in the ventral tegmental area (VTA) and substantia nigra

pars compacta (SNc) exhibit patterns of activity resembling the temporal-difference prediction error term in TDRL algorithms described above (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997). Specifically, when an animal receives an unexpected reward, these neurons rapidly and transiently increase their firing rate above their tonic, baseline firing rate. The magnitude of this phasic burst of activity scales with changes in expected value in a manner that reflects a quantitative prediction-error signal: the response correlates positively with reward magnitude and negatively with the expectation of receiving reward; receipt of a fully predicted reward evokes minimal response; and when an expected reward is omitted, dopamine neurons briefly pause firing at the time when reward was expected (Fiorillo et al., 2003; Morris et al., 2004; Bayer and Glimcher, 2005; Tobler et al., 2005). Although dopamine neurons' low tonic firing rate effectively provides a floor that limits the dynamic range available for signaling negative prediction errors through reductions in firing rate, the duration of this pause has been shown to encode negative prediction errors (Bayer et al., 2007), and symmetrical changes in terminal dopamine release have been observed for positive and negative prediction errors (Hart et al., 2014).

Also essential to the notion that phasic dopamine responses signal temporal-difference prediction errors is the observation that the unexpected presentation of a reward-predictive stimulus evokes burst firing that scales in magnitude with the expected value associated with that stimulus (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005). That is, a temporal-difference prediction error is evoked by any change in expectation of future reward (equation 1.4), such as that caused by the transition into a reward-predictive state denoted by a conditioned stimulus even if no actual reward is delivered at that

moment. This phasic dopamine response to a reward-predictive cue at a moment when no actual reward was expected or received therefore provides a readout of the associated cached value, an experimental approach I used to examine dopamine-associated cached values (Chapters 2 and 3). Indeed, many researchers have employed this approach, finding that cue-evoked dopamine responses (and thus the associated cached values) incorporate not only the objective expected value (reward magnitude times probability) (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Gan et al., 2010; Pasquereau and Turner, 2013), but also several more subjective factors such as idiosyncratic flavor and risk preferences (Nasrallah et al., 2011; Sugam et al., 2012; Lak et al., 2014; Stauffer et al., 2014), preference for sooner rather than delayed reward (Kobayashi and Schultz, 2008; Day et al., 2010), and even preference for advanced information that does not affect the likelihood of receiving reward (Bromberg-Martin and Hikosaka, 2009). These observations that subjective attributes affecting animals behavioral preferences are incorporated into the cached values reported by cue-evoked dopaminergic prediction-error signals are consistent with the notion that this signal is encoded in a common currency of subjective value (utility) such that these cached values could provide the basis for action selection (Kable and Glimcher, 2009; Lee et al., 2012; Schultz, 2013).

It is worth clarifying, however, that neither I nor most others in this field are claiming that the cue-evoked dopamine response itself is used for determining the decision. This signal, by virtue of encoding temporal-difference prediction errors, is simply providing experimenters with a readout that correlates with the cached value of the option denoted by that cue. The purported function of dopaminergic prediction errors is as a teaching signal critical for learning the subjective value of stimuli and actions (Kable and Glimcher, 2009; Schultz, 2013); this teaching

function is theorized to be the same regardless of whether the dopaminergic prediction-error signal was evoked by unexpected reward or reward-predictive cue presentation. The cached values themselves are thought to be stored as synaptic weights at corticostriatal synapses (Daw and Doya, 2006; Cohen and Frank, 2009; Kable and Glimcher, 2009; Glimcher, 2011; Hong and Hikosaka, 2011; Lee et al., 2012), with different ensembles of striatal GABAergic medium spiny neurons (MSNs) representing possible actions via specific "output channels" through the basal ganglia (Graybiel, 1995; Mink, 1996; Redgrave et al., 1999; Hikosaka et al., 2000). Indeed, the activity of many MSNs throughout the striatum has been found to encode value-related signals thought to reflect the cached values associated with different reward-seeking actions (Samejima et al., 2005; Lau and Glimcher, 2008; Kim et al., 2009; Roesch et al., 2009; van der Meer and Redish, 2009; Cai et al., 2011; Seo et al., 2012).

The teaching function subserved by dopaminergic prediction-error signals is to modulate the synaptic weights onto striatal MSNs from presynaptic cortical and limbic inputs thought to represent the current state and potential actions from that state. More abstract accounts of this learning process (Daw and Doya, 2006; Kable and Glimcher, 2009; Glimcher, 2011; Lee et al., 2012) simply describe the strengthening of these corticostriatal synapses by positive prediction errors and their weakening by negative prediction errors through mechanisms supporting long-term potentiation (LTP) and long-term depression (LTD), respectively (Wickens et al., 1996; Reynolds et al., 2001). Others (Cohen and Frank, 2009; Hong and Hikosaka, 2011; Morita et al., 2013) have proposed elaborated models constrained by more detailed basal ganglia anatomy and physiology, such as the differential expression of D1 and D2 dopamine receptors on MSNs of the direct and indirect pathways, respectively, and dopamine's more selective role in mediating LTP

and LTD at corticostriatal synapses onto each respective pathway (Surmeier et al., 2007; Kreitzer and Malenka, 2008; Shen et al., 2008; Gerfen and Surmeier, 2011). In models at either level of detail (and specifically in the direct pathway for the latter more biologically-constrained class), when cortical inputs representing a given state or state-action pair coincide with the selection of a particular action through its basal ganglia output channel, dopamine released in response to a received reward that is greater than expected will strengthen only those active corticostriatal synapses, whereas a transient decrease in dopamine tone following rewards that are less than expected leads to the weakening of these synapses. Reciprocal effects occur at active synapses onto indirect pathway MSNs included in the biologically constrained models (Cohen and Frank, 2009; Hong and Hikosaka, 2011), with positive prediction errors supporting LTD and negative prediction errors supporting LTP at these synapses. Although some unresolved timing considerations remain (Izhikevich, 2007; Pawlak et al., 2010), particularly regarding the neurobiological implementation of so-called "eligibility traces" that at the algorithmic level facilitate the assignment of credit for obtained reward to the appropriate preceding actions (Sutton and Barto, 1998), there is at least some empirical evidence for striatal representations of selected actions persisting beyond the time of reward delivery (Kim et al., 2007; Lau and Glimcher, 2007; Kim et al., 2009).

The next question is how this dopamine-mediated learning of state-action values at corticostriatal synapses ultimately influences action selection. This process entails a winner-take-all mechanism for selecting one action to execute out of the set of all possible actions available in a given state. This filtering from many possible actions to the selection of one winner is precisely a major proposed function of the basal ganglia (Graybiel, 1995; Mink, 1996; Redgrave et al., 1999;

Hikosaka et al., 2000). One of the most complete descriptions currently available for such a process comes from the work of Hikosaka and others investigating decisions enacted through the oculomotor system in nonhuman primates (Hikosaka et al., 2000). Rapid, ballistic eye movements (saccades) are triggered from the superior colliculus (SC), which contains a topographic map of egocentric visual space and receives tonic inhibition from a major output nucleus of the basal ganglia, the GABAergic substantia nigra pars reticulata (SNr). Activation of direct pathway MSNs associated with a particular saccade direction inhibit corresponding SNr neuron activity, disinhibiting neurons in the corresponding part of SC that in turn trigger a motor command for that saccade. Prior dopamine-mediated learning of state-action values at specific corticostriatal synapses as described above results in cortical state representations eliciting the preferential activation of MSNs for a given saccade in that particular state. In contrast, weaker corticostriatal synapses associated with lower state-action values would be less likely to become activated, meaning that SC neurons for alternative saccade directions would remain suppressed by SNr inhibition and therefore would be less likely to be selected. While this particular work focuses on the generation of saccadic eye movements, other types of decisions from arm movements and locomotion to more abstract cognitive planning also are thought to involve analogous pathways through the basal ganglia to outputs for their respective effectors in brainstem motor nuclei or via thalamocortical projections.

An open question remains regarding the extent to which value-guided action selection is determined within the basal ganglia and conveyed via outputs to these brainstem motor nuclei versus instead providing a strong biasing signal influencing selection that ultimately occurs within the cortex. Indeed, some neuroeconomics researchers posit an ultimately cortical locus for

choice, but one that is biased by feedback from the basal ganglia relayed through the thalamus and intracortical connectivity (Kable and Glimcher, 2009; Lee et al., 2012). For example, the lateral intraparietal cortex (LIP) has been extensively investigated as a site where putative representations of the value of different saccade directions might be compared for the selection of a winning action direction, communicated to the SC via LIP's interconnections with the frontal eye fields. Neural correlates of a variety of value-related parameters have been recorded in numerous locations throughout the frontal and parietal cortices (Platt and Glimcher, 1999; Dorris and Glimcher, 2004; Sugrue et al., 2004; Padoa-Schioppa and Assad, 2006; Kim et al., 2008; Louie and Glimcher, 2010; Kennerley et al., 2011; Sul et al., 2011; Cai and Padoa-Schioppa, 2014; Strait et al., 2014). A sub-focus of current research within this field, related to the question of which neural structures are most critical for making value-guided choices, pertains to the level of abstraction at which animals are comparing between and ultimately deciding based upon these subjective values. Specifically, are choices are made in "goods space," using orbitofrontal cortex (OFC) representations of expected outcome values independent of the particular action required to obtain them (Padoa-Schioppa and Assad, 2006; Wunderlich et al., 2010; Hare et al., 2011; Padoa-Schioppa, 2011; Cai and Padoa-Schioppa, 2014), or do choices instead depend on selection within "action space," using some combination of frontoparietal and/or corticostriatal circuitry wherein value representations are specifically associated with the motoric elements required for selecting these options and obtaining their outcomes (Platt and Glimcher, 1999; Redgrave et al., 1999; Dorris and Glimcher, 2004; Glimcher et al., 2005; Samejima et al., 2005; Lau and Glimcher, 2007, 2008; Louie and Glimcher, 2010; Seo et al., 2012)? Because the latter framework still is the more commonly described within this field, I primarily have introduced my doctoral work through the lens of

"action selection" by using language in accord with this framework. As I discuss in my final chapter, however, the results of my experiments actually may lend greater support for this dopamine-related learning system's involvement in valuation of "goods" when defined more restrictively as outcomes associated with stimuli with which separate representations of response costs still must be combined prior to their use for action selection (Rangel and Hare, 2010; Rangel and Clithero, 2014).

In sum, the current prevailing theories of the neural basis of value-guided decision making converge on the position that cached values for possible actions leading to reward from a given state are compared for determining choices. These cached values are thought to be stored within corticostriatal synapses with weights that are learned and updated according to dopaminergic prediction-error signals. A property of this TDRL-like prediction-error system is that the cached value of a given state-action pair can be read out from the dopamine response to the unexpected presentation of the reward-predictive cue denoting that state. Because these cached values are theorized to serve as the sole basis for selecting actions, the rank ordering of the cached values reported by cue-evoked dopamine responses to different available options should, by definition, align with the ordinal utility of these options. However, there was some initial evidence conflicting with aspects of these theories that informed my doctoral experiments that ultimately challenged this general premise.

1.2) Initial inconsistencies with the prevailing theories

The general premise that action selection is based upon cached values learned through a dopamine-dependent TDRL-like learning system hinges upon the alignment of these dopamine-associated cached values with animals' subjective preferences as revealed through their choice behavior. Namely, for these cached values to serve as the sole basis of choice, they need to incorporate any and all factors that affect animals' preference and must do so exactly to the extent that these factors influence preference. Indeed, as described in the sections above, there have been numerous reports in which these cached values, as read out from cue-evoked dopamine responses, do incorporate a variety of economic attributes affecting animals' preferences (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Roesch et al., 2007; Kobayashi and Schultz, 2008; Bromberg-Martin and Hikosaka, 2009; Day et al., 2010; Nasrallah et al., 2011; Sugam et al., 2012; Lak et al., 2014). Whereas most of these studies manipulated one dimension at a time, a recent study found that these dopamine-associated cached values integrate animals' flavor and risk preferences along with objective expected value into a common currency of subjective value (Lak et al., 2014). This correspondence between the cached values recorded from phasic dopamine responses and the utility functions measured from the animals' behavior provided some of the most detailed and compelling evidence consistent with the proposal that these cached values provide sufficient information about the subjective value of possible actions for animals to make their decisions based upon these cached values alone.

No matter how many positive correlations are observed between dopamine-associated cached values and behavioral preferences, however, the existence of counterexamples would demonstrate that this fundamental claim of decision-making theories cannot hold as a general principle. Around the time that I joined the Phillips lab, Jerylin Gan, Mark Walton, and Paul Phillips published their findings that cue-evoked dopamine release did not consistently encode effortful response costs (Gan et al., 2010). Specifically, in their cost manipulation, each option yielded one food pellet reward but required a different number of lever presses to obtain it: 16 presses versus 2 presses in one session type, 16 versus 32 presses in another session type. Although they did initially observe greater cue-evoked dopamine release to the low-effort option in the former condition, this difference was not observed in a second recorded session nor in a subsequent group receiving at least nine sessions of training with these contingencies. Likewise, there was no significant difference in dopamine release to the 16- versus 32-press options at any time point recorded, despite robust behavioral preferences for the low-effort option in all variants of this effort manipulation. In stark contrast, in sessions that manipulated the number of pellets earned (4 vs. 1 or 1 vs. 0 pellets) for the same effort requirement (16 presses), they observed significantly greater cue-evoked dopamine release for the option yielding the larger reward size across all variants of this reward manipulation. Importantly, they also found that, on average, the rats exhibited behavioral indifference between the high value option from the reward manipulation and the low effort option from the cost manipulation when tested head-to-head in a separate behavioral session, demonstrating that each manipulation conferred comparable differences in utility between the options within their respective session types. Although another group has since reported that effortful response costs are encoded to some extent by cue-evoked dopamine release (Day et al., 2010), their finding has not replicated in our hands even when

using initial training conditions more similar to their experimental design (Arnold et al., unpublished). Although data collection is still ongoing for some groups within our latter experiment, across multiple groups of animals performing an effort manipulation we have clear results that dopamine-associated cached values do not incorporate anticipated effort to the extent that it robustly influences animals' behavioral preferences. Others in the Phillips lab also have found that cue-evoked dopamine release remains relatively constant despite escalating response costs in a progressive ration task (Wanat et al., 2010). Finally, two studies recording the activity of SNc dopamine neurons in rhesus macaques found few cells that reliably encoded anticipated effort requirements (Ravel and Richmond, 2006; Pasquereau and Turner, 2013). In particular, the latter study observed that although a minority (~11%) of dopamine neurons were sensitive to reward and effort in a manner seeming to reflect an integrated subjective-value signal, the majority of cue-responsive dopamine neurons only signaled expected reward but not effort, largely consistent with the conclusions of Gan et al. (2010).

This weak incorporation of anticipated effort into dopamine-associated cached values also may generalize to other forms of economic costs, such as the aversiveness associated with noxious or harmful stimuli. This question of how dopamine neurons respond to aversive stimuli certainly has been a longstanding source of controversy, and there have been multiple conflicting reports regarding whether aversive stimuli cause phasic increases or decreases in dopamine transmission or even reveal heterogeneous subpopulations of dopamine neurons exhibiting each type of response (Mirenowicz and Schultz, 1996; Horvitz, 2000; Ungless et al., 2004; Joshua et al., 2008; Brischoux et al., 2009; Matsumoto and Hikosaka, 2009; Bromberg-Martin et al., 2010a; Wang and Tsien, 2011; Cohen et al., 2012; Oleson et al., 2012; Fiorillo, 2013; Fiorillo et al.,

2013b; Fiorillo et al., 2013a). The most recent of these publications (Fiorillo, 2013) has argued that many of these previous studies did not adequately demonstrate how aversive their various noxious stimuli actually were to the animals, making the negative subjective value of these stimuli difficult to compare to the positive subjective value of appetitive stimuli presented within some of these same studies. Using behavioral choice procedures to directly determine the subjective value of rewarding and aversive stimuli presented to his animals, Fiorillo (2013) concluded that dopamine neurons are relatively insensitive to the aversiveness signaled by these conditioned stimuli. From these experiments, he also found that although different dopamine neurons did exhibit some variability in their responses to aversive stimuli, these responses seemed to fall along a single continuum (Fiorillo et al., 2013a) rather than segregating into discrete populations of responses as previously suggested (Matsumoto and Hikosaka, 2009). Although characterizing the extent of diversity of dopamine neuron activity and functions remains an important and active area of ongoing research (Bromberg-Martin et al., 2010a; Lammel et al., 2012; Chaudhury et al., 2013), the broader conclusions from Fiorillo's recent work converge with those drawn from the studies using effort manipulations (Ravel and Richmond, 2006; Gan et al., 2010; Wanat et al., 2010; Pasquereau and Turner, 2013) supporting the notion that certain cost-related economic attributes are not reliably incorporated into dopamine-associated cached values to the extent that they influence behavioral preferences.

This differential incorporation of certain economic attributes into dopamine-associated cached values call into question whether these cached values represent subjective value in a truly domain-general manner, and therefore could have quite profound implications for neuroeconomic theories of decision making that depend on the consistent ordering of these

cached values relative to observed behavioral preferences. For instance, when decisions involve tradeoffs between options differing along dimensions that are weakly versus strongly incorporated into the cached values, there might be specific circumstances in which animals most often choose an option that does not have the greatest dopamine-reported cached value, which would violate the fundamental premise of current theories of decision making. This is precisely what we tested and ultimately observed in the experiments presented in Chapter 2, using a mixed-contingency decision-making task in which we manipulated both the reward value and effortful response cost associated with each option (Hollon et al., 2014). Although we predicted that we would find this result based on our previous work involving independent manipulations of reward and effort (Gan et al., 2010), and while similar conclusions regarding the mismatch between cached values and the subjective value of actions could be inferred from related work (Fiorillo, 2013; Pasquereau and Turner, 2013), to my knowledge the results presented in Chapter 2 are the first demonstrating that animals will exhibit a preference for an option associated with lower dopamine release than that of a non-preferred alternative. The task employed by Pasquereau and Turner (2013) during their dopamine neuron recordings also included presentation of individual stimuli with mixed reward and effort contingencies, but their behavioral choice task only included choices between options that differed along a single dimension. Had they included choices between their low-value / low-effort and high-value / high-effort options, then it might have been possible to discern whether the larger dopamine neuron response they observed to the latter option also provided evidence for an inversion between the rank ordering of dopamine-associated cached values and subjective preferences as we observed in our experiments. Likewise, Fiorillo's behavioral experiments showed that the monkeys demonstrated a significant preference for a small-reward option over an option

associated with larger reward plus aversive air puff (Fiorillo et al., 2013b), but this series of papers did not present the data for recordings of dopamine neuron activity in response to the small-reward cue. Comparing these responses to that for the large-reward-plus-air-puff cue (Fiorillo, 2013) would have permitted the examination of whether the non-preferred option associated with a larger reward still evokes a greater dopamine response than the preferred small-reward option, as I observed in my experiments. In the concluding chapter of this dissertation, I expand upon our discussion of these results from Chapter 2 to elaborate on both the more nuanced role that dopamine and these associated cached values might play in value-guided decision making as well as the possible neural substrates providing the additional information required if these cached values are to contribute to these decisions.

1.3) Addressing an unresolved discrepancy

As discussed above (section *1.1c*), the prominent neuroeconomic theories discussed throughout do not posit that the cue-evoked dopamine response itself is used as a valuation signal for guiding immediate action selection. Rather the purpose of this signal would simply be to teach an animal whether some even earlier state or action led to the presentation of that cue. In experimental designs in which cue presentation is not contingent on the animal's actions, these cue-evoked dopamine signals may well be effectively inconsequential or epiphenomal for the animal. Nevertheless, these cue-evoked dopamine signals provide a neurobiological readout of the associated cached values represented at the time of cue onset. This experimental approach for obtaining a neural proxy of the cached value has been widely used in previous work and is my

primary approach throughout the experiments in Chapters 2 and 3. Unlike the purported encoding of state-action values within the striatum and various cortical structures wherein these values for different actions could be compared for the determination of action selection (Platt and Glimcher, 1999; Dorris and Glimcher, 2004; Samejima et al., 2005; Lau and Glimcher, 2008; Louie and Glimcher, 2010; Cai et al., 2011; Seo et al., 2012), dopamine neurons do not seem to provide multiple, simultaneous valuation signals corresponding to different options available (Schultz, 1998; Morris et al., 2006; Roesch et al., 2007). The question of which cached value is reported during actual choice scenarios, when animals are deciding between concurrently available options, nevertheless provides important information for determining which variant of TDRL algorithms is enacted by this system; this question, however, has received far less attention compared to the multitude of studies recording dopamine transmission when only one option is available at a time. The few studies addressing this question arrived at conflicting conclusions regarding whether cue-evoked dopamine during choices represents the cached value of the subsequently chosen option (Morris et al., 2006) or the greatest cached value available regardless of which option is chosen (Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012), with the former supporting the neural instantiation of a SARSA-like TDRL algorithm (Rummery and Niranjan, 1994; Niv et al., 2006) whereas the latter is consistent with Q-learning (Watkins, 1989; Daw, 2007).

The divergent findings of these sets of studies frequently have been attributed to differences in species or recording location (Daw, 2007; Roesch et al., 2007; Morris et al., 2010; Morita et al., 2013; Morita, 2014; Morita and Kato, 2014). Morris et al. (2006) recorded SNc dopamine neurons in macaque monkeys whereas Roesch et al. (2007) recorded VTA dopamine neurons in

rats, and these latter findings were corroborated by recordings of dopamine release in the nucleus accumbens (Day et al., 2010; Sugam et al., 2012). However, there are reasons to question how conclusively the latter set of findings (Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012) actually implies the instantiation of a Q-learning algorithm. Specifically, the only point of disagreement between the results of this set of studies versus those of Morris et al. (2006) concerned the relative magnitude of dopamine transmission in trials where animals chose the non-preferred option with a lower cached value than that of the unchosen (but preferred, on average) alternative, but there are alternative explanations for why the studies reporting relatively homogenous dopamine responses across choices may have observed this result. It is possible that these choices for the option that was not preferred on average were mistakes, and the higher dopamine release observed in these trials (comparable to that in both choice and forced trials for the preferred option) may have reflected that the animals actually expected their preferred option in these trials, and the resulting dopamine response therefore may have misrepresented its cached value. Alternatively, it is also possible that these choices for the non-preferred option may have been deliberate exploratory decision instead, and the resultant dopamine response may have reflected an additional "exploration bonus" theorized to augment the dopamine response in these trials (Dayan and Sejnowski, 1996; Kakade and Dayan, 2002), plausibly to a level comparable to that in trials for the preferred option.

Although these alternative explanations would themselves be difficult to resolve, the findings reported in Chapter 2 actually present an opportunity to address our primary question before even needing to analyze the dopamine response during these anomalous choices for a non-preferred option. Specifically, the SARSA and Q-learning algorithms can be differentiated by

examining the prediction error signaled when animals choose an option that does not have the greatest cached value (equations 1.8 and 1.9), and we demonstrate in Chapter 2 that there are circumstances where animals exhibit behavioral preferences for an option that does not have the greatest cached value (Hollon et al., 2014). Therefore, the dissociation between subjective value of available options and their dopamine-associated cached values we observe in Chapter 2 provides an opportunity to analyze the dopamine response during choices for an option that does not have the greatest cached value yet is still the preferred option. By comparing the dopamine response in these trials to those in the single-option forced trials, we can determine which TDRL algorithm is being enacted without needing to base our conclusions on anomalous choices for a non-preferred option. This is precisely the approach we take to address this unresolved discrepancy in Chapter 3.

1.4) Measuring rapid changes in mesolimbic dopamine concentration

Throughout all experiments described in this dissertation, I used fast-scan cyclic voltammetry (FSCV) at chronically implanted carbon-fiber microelectrodes (Clark et al., 2010) to record rapid changes in extracellular dopamine concentration in the nucleus accumbens core of rats performing cost-benefit decision-making tasks (Hollon et al., 2014). The small diameter of the carbon fiber (7 μm) and the biocompatible polyimide-coated fused silica (90 μm) encasing it cause minimal detectable damage, permitting chronic implantation for multiple recordings from the same recording site in a given animal (Clark et al., 2010). FSCV permits detection of sub-second changes in extracellular dopamine concentration with an effective temporal resolution

(10 Hz) far superior to that of other sampling techniques such as microdialysis. Although throughout this dissertation I discuss fairly interchangeably results regarding dopamine transmission, whether obtained by measuring dopamine release with FSCV or recording putative dopamine neuron activity with extracellular electrophysiology, these complementary approaches of course are not the same and each has distinct advantages and disadvantages. For example, electrophysiological unit recordings provide even higher temporal resolution than does FSCV but come with the challenges of finding and definitively identifying cells as dopaminergic, whereas FSCV provides high chemical selectivity of actual dopamine release from a population of release sites near the electrode and will therefore also reflect any terminal regulation.

Additionally, I have largely uniformly discussed previous results regarding dopamine transmission across nigrostriatal and mesolimbic pathways, despite awareness of growing emphasis in the field regarding potential pathway-specific heterogeneity (Bromberg-Martin et al., 2010a; Lammel et al., 2012; Chaudhury et al., 2013) and dissociable contributions of striatal subregions to specific forms of reward-seeking behavior (Yin et al., 2008). All recordings in the experiments of my dissertation were restricted to the nucleus accumbens core. Although this may limit the generalizability of these results, a convergence of anatomical (Mogenson et al., 1980; Sesack and Grace, 2010), behavioral pharmacology (Salamone et al., 2007; Floresco et al., 2008; Nicola, 2010), neuroimaging (Bartra et al., 2013; Garrison et al., 2013), electrophysiological (Roesch et al., 2007) and neurochemical evidence (Phillips et al., 2003; Gan et al., 2010) provides a strong rationale for investigating dopamine transmission in the nucleus accumbens in particular for gaining a better understanding of value-guided decision making. Nevertheless, characterizing dopamine release within the dorsal striatum of behaving animals remains an

important area for future research, and although behavioral-event-related phasic release has seemed more difficult to detect in the dorsolateral striatum (Zhang et al., 2009; Brown et al., 2011; Howe et al., 2013), this may be better achieved by combining these chronically implantable electrodes with microdrive technology (Howe et al., 2013).

Chapter 2

Dopamine-associated cached values are not sufficient as the basis for action selection*

* This chapter was originally published with minor reformatting as an article in the *Proceedings of the National Academy of Sciences* with the same title and Monica M. Arnold, Jerylin O. Gan, Mark E. Walton, and Paul E. M. Phillips as coauthors. The full citation is as follows:

Hollon NG, Arnold MM, Gan JO, Walton, ME, Phillips PEM (2014) Dopamine-associated cached values are not sufficient as the basis for action selection. Proc Natl Acad Sci USA 111:18357-18362.

All authors contributed to experimental design. I collected the experimental data with assistance from M.M.A. and J.O.G., I carried out data analysis with assistance from M.M.A., and I prepared the manuscript with M.M.A. and P.E.M.P.

Abstract

Phasic dopamine transmission is posited to act as a critical teaching signal that updates the stored (or "cached") values assigned to reward-predictive stimuli and actions. It is widely hypothesized that these cached values determine the selection between multiple courses of action, a premise that has provided a foundation for contemporary theories of decision making. In the current work we used fast-scan cyclic voltammetry to probe dopamine-associated cached values from cue-evoked dopamine release in the nucleus accumbens of rats performing cost-benefit decision-making paradigms to critically evaluate the relationship between dopamine-associated cached values and preferences. By manipulating the amount of effort required to obtain rewards of different sizes, we were able to bias rats toward preferring an option yielding a high-value reward in some sessions and toward instead preferring an option yielding a low-value reward in others. Therefore, this approach permitted the investigation of dopamine-associated cached values in a context where reward magnitude and subjective preference were dissociated. We observed greater cue-evoked mesolimbic dopamine release to options yielding the high-value reward even when rats preferred the option yielding the low-value reward. This result identifies a clear mismatch between the ordinal utility of the available options and the rank ordering of their cached values, thereby providing robust evidence that dopamine-associated cached values cannot be the sole determinant of choices in simple economic decision making.

Introduction

In contemporary theories of economic decision making, values are assigned to reward-predictive states in which animals can take action to obtain rewards, and these state-action values are stored ("cached") for the purpose of guiding future choices based upon their rank order (Sutton and Barto, 1998; Daw and Doya, 2006; Rangel et al., 2008; Kable and Glimcher, 2009; Lee et al., 2012). It is believed that these cached values are represented as synaptic weights within corticostriatal circuitry, reflected in the activity of subpopulations of striatal projection neurons (Samejima et al., 2005; Lau and Glimcher, 2008; Cai et al., 2011; Tai et al., 2012), and updated by dopamine-dependent synaptic plasticity (Reynolds et al., 2001; Pawlak and Kerr, 2008; Shen et al., 2008). Indeed, there is a wealth of evidence suggesting that the phasic activity of dopamine neurons reports instances when current reward or expectation of future reward differs from current expectations (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Waelti et al., 2001; Bayer and Glimcher, 2005; Roesch et al., 2007; Glimcher, 2011; Cohen et al., 2012; Schultz, 2013; Steinberg et al., 2013; Hart et al., 2014; Stopper et al., 2014). This pattern of activity resembles the prediction-error term from temporal-difference reinforcement-learning algorithms, which is considered the critical teaching signal for updating cached values. A notable feature of models that integrate dopamine transmission into this computational framework is that the cached value of an action is explicitly read out by the phasic dopamine response to unexpected presentation of a cue that designates the transition into a state in which that action yields reward. Therefore, cue-evoked dopamine signaling provides a neural representation of the cached values of available actions, and if these cached values serve as the basis for action

selection, then cue-evoked dopamine responses should be rank ordered in a manner that is consistent with animals' behavioral preferences.

Numerous studies that recorded cue-evoked dopamine signaling have reported correlations with the expected utility (subjective value) of actions (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Roesch et al., 2007; Kobayashi and Schultz, 2008; Bromberg-Martin and Hikosaka, 2009; Day et al., 2010; Gan et al., 2010; Nasrallah et al., 2011; Sugam et al., 2012; Pasquereau and Turner, 2013; Lak et al., 2014; Stauffer et al., 2014). For example, risk-preferring rats demonstrated greater cue-evoked dopamine release for a risky option than for a certain option of equivalent objective expected value (reward magnitude times probability), whereas risk-averse rats showed greater dopamine release for the certain than for the risky option (Sugam et al., 2012). Likewise, the cached values reported by dopamine neurons in macaque monkeys accounted for individual monkeys' subjective flavor and risk preferences, with each attribute weighted according to its influence on behavioral preferences (Lak et al., 2014; Stauffer et al., 2014). These observations, consistent across measures of dopamine neuronal activity and dopamine release, reinforce the prevailing notion that the dopamine-associated cached values could be the primary determinant of decision making (Daw and Doya, 2006; Kobayashi and Schultz, 2008; Rangel et al., 2008; Kable and Glimcher, 2009; Day et al., 2010; Sugam et al., 2012; Lee et al., 2012; Schultz, 2013; Lak et al., 2014; Stauffer et al., 2014) because the cue-evoked dopamine responses were rank ordered according to the animals' subjective preferences. However, there have been some reports that other economic attributes such as effortful response costs (Ravel and Richmond, 2006; Gan et al., 2010; Wanat et al., 2010; Pasquereau and Turner, 2013) or the overt aversiveness of an outcome (Fiorillo, 2013) are inconsistently represented by

cue-evoked dopamine responses. For example, Gan et al. (Gan et al., 2010) showed that independent manipulations of two different dimensions (reward magnitude and effort) which had equivalent effects on behavior did not have equivalent effects on dopamine release. Paralleling these findings, a recent report reached a similar conclusion that dopamine transmission preferentially encodes an appetitive dimension but is relatively insensitive to aversiveness (Fiorillo, 2013).

Because these cue-evoked dopamine signals represent cached values that are purported to determine action selection, their differential encoding of economic dimensions has potentially problematic implications in the context of decision-making. Namely, by extrapolating from these studies (Ravel and Richmond, 2006; Gan et al., 2010; Wanat et al., 2010; Fiorillo, 2013; Pasquereau and Turner, 2013), one might infer that when a decision involves the tradeoff between these economic dimensions, the rank order of the dopamine-associated cached value for each of the available options would not consistently reflect the ordinal utility of these options and therefore could not on their own be the basis of choices. However, this counterintuitive prediction was not explicitly tested by any of these previous studies, and thus it remains a provocative notion that merits direct examination, as it is contrary to the prevailing hypothesis described above which is fundamental to contemporary theories of decision making. Therefore, we investigated interactions between dimensions that have previously been shown during independent manipulations to be weakly or strongly incorporated into these cached values. Specifically, we increased the amount of effort required to obtain a large reward such that animals instead preferred a low-effort option yielding a smaller reward, and we used fast-scan cyclic voltammetry to record cue-evoked mesolimbic dopamine release as a neurochemical

proxy for each option's cached value. These conditions permitted us to test whether or not the cached values reported via cue-evoked dopamine indeed align with animals' subjective preferences across these mixed cost-benefit attributes.

Methods

Subjects and Surgery. All procedures were approved by the University of Washington Institutional Animal Care and Use Committee. A total of 41 male Sprague Dawley rats (Charles River Laboratories), 250-300 grams upon arrival, were used for this study. Fourteen rats in the first cohort and 14 in the second cohort completed recording sessions included in this study; four additional rats were excluded due to electrode misplacement, three due to failure of electrodes to satisfy criteria for dopamine detection, and six due to post-surgical complications (e.g., head-cap loss). Rats were maintained on a 12-hour light/dark cycle (lights on at 0700), with all behavioral testing occurring during the light phase. Rats were pair-housed until surgery, after which they were housed individually. Rats were anesthetized with isoflurane for bilateral implantation of carbon-fiber microelectrodes (Clark et al., 2010) targeting the nucleus accumbens core (1.3 mm anterior, 1.3 mm lateral, 6.8-7.0 mm ventral to bregma; Figure 2.1) and a Ag/AgCl reference electrode. After at least one week recovery post-surgery, rats were food-restricted to 90% their *ad libitum* body weight; for all subsequent behavioral procedures, each rat received a total of ~15 grams of food per day consisting of pellets earned as reward during behavioral sessions plus standard lab chow after these sessions. Water was available *ad libitum* in the animals' home cages.

Initial Behavioral Training. In their homecages prior to the first session of training, the food-restricted rats were exposed to the 45-mg food pellets (Bio-Serv dustless precision pellets) that served as rewards for all subsequent sessions. All training sessions took place between 0800 and 1800 in one of four standard operant chambers (Med Associates, VT). Each chamber was equipped with a central food magazine and magazine light, a retractable lever on either side of the magazine (6 cm above the grid floor), a cue light above each lever, a house light at the top back left corner of the chamber, and a ventilation fan on the back wall of the sound-attenuating cabinet around the operant chamber.

Rats underwent one session of magazine training in which a total of 60 food pellets were delivered non-contingently, one at a time with a variable time interval (60 ± 20 s). Training to press levers for food pellets began the next day. One of the two cue lights was illuminated (side counterbalanced between rats), the corresponding lever was extended continuously for the duration of the session, and each press was reinforced on a fixed ratio (FR) 1 continuous reinforcement schedule until rats received 100 pellets in a two-hour session. If rats did not press the lever within 15-20 min, a pellet was placed behind the lever to encourage the rat to interact with the lever. In the next session, the other cue light was illuminated and lever extended for another 100-pellet session reinforced on a continuous FR-1 schedule.

All subsequent sessions consisted of training on discrete trials, such that after completing the response requirement, the cue light turned off, the lever retracted, a pellet was delivered into the food magazine, and the magazine light was illuminated for six seconds, followed by a variable

inter-trial interval (ITI). At this stage of training, only one lever was available on each trial (all trials were "Forced"), and each session consisted of 80 total trials (40 for each lever). For the first cohort of rats, each trial began at the end of the ITI with the simultaneous onset of a cue light and extension of the corresponding lever. For the second cohort, the cue light preceded the lever extension by 5s. Rats completed one session of FR-1 with a 20 ± 5 s ITI and unlimited time to initiate responding on each trial. For all subsequent behavioral sessions, rats were connected to a head-stage containing a voltammetric amplifier to habituate them to the equipment used for eventual recording sessions. While 'tethered,' subsequent training consisted of one session each of FR-1 with a 20 ± 5 s ITI, FR-4 with a 30 ± 10 s ITI, FR-8 and FR-16 with a 45 ± 15 s ITI for all sessions thereafter. Starting in the FR-16 session, failure to initiate responding within 10 s of lever presentation resulted in an unrewarded "Miss." Training on FR-16 sessions continued until rats completed over 90% of the trials in a session. Rats then performed behavioral decision-making sessions that included blocks of four single-option "Forced" and four dual-option "Choice" trials, during which either the reward magnitude or effort requirement differed between the two options, as in our previous study that included independent manipulations of reward and effort (Gan et al., 2010). For the reward manipulation sessions, each option required four lever presses, with one lever yielding four pellets and the other yielding one pellet; for the effort manipulation sessions, each option yielded one food pellet, with one requiring four lever presses and the other requiring 32 presses. The contingencies assigned to each lever side were reversed between each session, and each rat performed daily sessions of either the reward or effort manipulation (order counterbalanced) until it reached criterion (75% choice in a sliding window of 12 Choice trials) in fewer than 80 trials for both lever side assignments. After completing both

the reward and effort manipulation stages, rats then advanced to the mixed-contingency decision-making task described below.

Mixed-Contingency Decision-Making Task. All sessions consisted of blocks of four single-option "Forced" trials, in which only one of the two options was available, followed by four "Choice" trials, in which both options were concurrently available. In Choice trials, the unchosen lever retracted and cue light turned off once rats made an initial press on the chosen lever. A 45 ± 15 s variable inter-trial interval separated each discrete trial, with a maximum of 120 trials per session. As in prior training, for the first cohort of rats each trial began immediately after the ITI with the onset of one or both cue lights and the simultaneous extension of the corresponding lever(s), and for the second cohort, the lever(s) extended 5 s after the onset of the cue light(s).

The mixed-contingency decision-making task consisted of two types of sessions: "Moderate Cost" and "High Cost" conditions (Figure 2.2*A*). In both conditions, one lever served as a low-value/low-effort (LL) reference option, yielding one food pellet for four lever presses. The alternative option yielded a high-value reward (four pellets) for a medium effort requirement (eight presses) in the "Moderate Cost" condition (high-value/medium-effort: HM) or for a high effort requirement in the "High Cost" condition (high-value/high-effort: HH). Prior to conducting voltammetric recordings, this high effort requirement was determined individually for each rat such that they preferred the LL option. For the first cohort, the lever side assigned to the low- vs. high-value options were reversed every two behavioral sessions, and if a rat did not reliably prefer the LL option in both side configurations, the effort requirement for the HH option was increased by eight presses for the subsequent High Cost session, but always remained

constant within a given session. The same procedure was used to determine the high effort requirement for the second cohort, except the lever side assignments were reversed between every one to three sessions pseudo-randomly. The final high effort requirements used in recording sessions ranged from 32 to 48 lever presses between rats in the first cohort, 32-128 presses for rats in the second cohort. For both cohorts, recordings were conducted after rats had performed at least eight behavioral sessions of a given condition (Gan et al., 2010) and were always conducted on the second session with a given lever side assignment. Behavioral criterion was defined as 75% choice for the HM option in the Moderate Cost condition and for the LL option in the High Cost condition within a sliding 12-Choice window. After reaching this criterion, rats performed four additional blocks (32 trials), which provided the primary data analyzed from each recording session. We also obtained recordings from High Cost sessions in which rats did not reach the intended criterion for the LL option and instead reached the opposite criterion, preferring the HH option. The high effort requirements from these HH-preferred High Cost sessions ranged from 32-64 presses for the first cohort and 32-128 presses for the second cohort.

Fast-Scan Cyclic Voltammetry Recording Sessions. The chronically implanted carbon-fiber microelectrodes were connected to a head-mounted voltammetric amplifier for dopamine detection by fast-scan cyclic voltammetry as previously described (Clark et al., 2010). A potential of -0.4 V (versus the Ag/AgCl reference) was applied to the carbon fiber and ramped to +1.3 V and back at a rate of 400 V/s. This voltammetric scan was applied at a frequency of 60 Hz for ~40 min prior to recording the behavioral sessions and then at 10 Hz for ~20 min prior to and throughout the recording session. To confirm that electrodes were capable of detecting

chemically verified dopamine, a series of unexpected food pellets were delivered before and after each recording session. The voltammetry data from a recording session were included in the analysis only if the pre- and post-session pellet deliveries elicited dopamine release whose cyclic voltammogram (electrochemical signature) achieved a high correlation ($r^2 \geq 0.75$ by linear regression) with that of a dopamine standard.

Statistical Analyses. Post-criterion choice proportions were normalized with the arcsine transformation and compared to indifference using two-tailed, one-sample t-tests in SPSS. Voltammetry data analysis was carried out using software written in LabView and Matlab. Following 2000-Hz low-pass filtering, dopamine was isolated from the background-subtracted (one second prior to cue onset) voltammetric signal using chemometric analysis (Heien et al., 2005) using a standard training set based on stimulated dopamine release detected by chronically implanted electrodes (Clark et al., 2010). Dopamine concentration was estimated based on the average post-implantation electrode sensitivity. Noise spikes >1.5 nA versus the immediately preceding and following time points were removed (Gan et al., 2010), and the data were smoothed using a 0.5-s moving average.

The discriminability of cue-evoked dopamine responses in the different Forced trial types was analyzed at each time point using the area under the receiver operating characteristic curve (auROC), an approach from signal detection theory (Green and Swets, 1966). The high-value option (HM or HH) was always coded as the positive cases in comparisons with the LL option, and auROC values were *not* rectified around 0.5 (i.e., if the LL option had evoked a greater dopamine response, the auROC values would have been less than 0.5). Significant

discriminability at each time point was determined using a random permutation test, shuffling the trial types and re-computing the auROC, and repeating this for 2000 permutations to generate a null distribution. After correcting for multiple comparisons across time using a suprathreshold cluster-correction technique (Nichols and Holmes, 2002; Buschman et al., 2012), all time points outside the 95% confidence interval were considered statistically significant. For graphical display purposes, all auROC values were transformed to a dopamine discriminability index ranging from -1 to 1: discriminability index = 2 * (auROC – 0.5).

To test the relationship between dopamine-associated cached values and subjective preference, we computed a dopamine discriminability index and a choice index to summarize each recorded session. The mean change in dopamine concentration over 5 s following cue onset for each post-criterion trial was used to calculate the auROC for a given session (HM or HH trials as positive cases, LL trials as negative cases), and each session's auROC was transformed to a dopamine discriminability index as above: discriminability index = 2 * (auROC – 0.5). Thus, a dopamine discriminability index approaching one indicates a greater dopamine response to the high-value option (HM or HH), whereas an index of -1 indicates a greater response to the LL option, and an index of 0 indicates equivalent dopamine release to either option. Likewise, the choice index was calculated by transforming the post-criterion choice behavior in each session: choice index = 2 * (p(H) – 0.5), where p(H) is the proportion of choices for the high-value option (HM or HH), such that a choice index of 1 corresponds to 100% choice for the high-value option, -1 indicates 100% choice for the LL option, and 0 indicates indifference between the two options. Categorical models of expected utility and expected benefits were evaluated with binomial tests of the number of sessions violating or satisfying the models' predictions, and regression models

were evaluated by comparing the goodness-of-fit using the second-order bias-corrected Akaike Information Criterion (AICc) (Akaike, 1974; Burnham and Anderson, 2002) based on the residual sum of squares.

Finally, to test for the possibility of direction-selective encoding by mesolimbic dopamine, we examined the counterbalanced pairs of recorded sessions from the first cohort of rats to compare the cue-evoked dopamine response during Forced trials for each option when it was assigned to the lever side ipsilateral versus contralateral to the hemisphere of the recording electrode. Within each trial type, the dopamine responses were indistinguishable between the two lever side assignments (Figure 2.3). Moreover, we observed the same pattern of greater dopamine transmission for the high-value option regardless of the lever assignment configuration. Because these results do not reflect direction encoding by mesolimbic dopamine, for the second cohort we included all recorded sessions meeting the electrochemical and behavioral criteria regardless of whether the counterbalanced pair was obtained.

Histological Verification of Recording Site. Animals were anesthetized with ketamine (100 mg/kg) and xylazine (20 mg/kg), and the recording site was marked by passing a current (~70 μA) through the carbon-fiber microelectrode for 20 s to make a small electrolytic lesion. Animals were perfused transcardially with physiological saline and then with four-percent paraformaldehyde in phosphate-buffered saline, in which brains also were post-fixed following removal from the skull. Brains were sunk in 15% sucrose solution in PBS for 24 hours, 30% sucrose for at least 72 hours, flash frozen in dry ice, sectioned coronally (30-60 μm) on a cryostat, mounted on slides, and stained with a 0.5% cresyl violet solution.

Results

Food-restricted rats performed an instrumental decision-making task with mixed reward and effort contingencies (Figure 2.2*A*). Sessions consisted of repeating blocks of four single-option "Forced" trials, in which only one of the two options was available, followed by four "Choice" trials, in which both options were concurrently available. After a 45 ± 15 s variable inter-trial interval, each trial began with the onset of one or both cue lights and the simultaneous extension of the corresponding lever(s). One lever was a low-value/low-effort (LL) reference option, yielding one food pellet for four lever presses. The alternative option yielded a high-value reward (four pellets) for a medium effort requirement (eight presses) in the "Moderate Cost" condition (high-value/medium-effort: HM) or for a high effort requirement in the "High Cost" condition (high-value/high-effort: HH). During initial training prior to recording dopamine transmission in the High Cost condition, this high effort requirement was determined individually for each rat such that they preferred the LL option. This high response cost ranged from 32 to 48 lever presses between rats but remained constant within each session for a given rat. A pair of voltammetry recordings (one session per counterbalanced lever side assignments for the high- and low-value options) was conducted for each condition. Behavioral criterion was defined as 75% choice for the HM option in Moderate Cost sessions and for the LL option in High Cost sessions within a sliding window of 12 consecutive Choice trials. After reaching this criterion, rats performed four additional blocks (32 trials), which provided the data analyzed from each session. Rats' post-criterion choices revealed significant preferences for the HM

option in the Moderate Cost condition (Figure 2.2*B*; $t_8 = 7.095$, $P = 0.0001$) and for the LL option in the High Cost condition (Figure 2.2*C*; $t_6 = 2.923$, $P = 0.0265$).

To monitor mesolimbic dopamine transmission, we used fast-scan cyclic voltammetry at carbon-fiber microelectrodes chronically implanted in the nucleus accumbens core (Clark et al., 2010) (Figure 2.1). Voltammetric recordings from post-criterion trials revealed phasic increases in dopamine concentration following cue onset for all trial types. Cue-evoked dopamine release during Forced trials in the Moderate Cost condition was greater in HM trials than in LL trials (Figure 2.2*D*). To quantify this selectivity, at each time point we calculated a dopamine discriminability index (Kepecs et al., 2008) based on the area under the receiver operating characteristic (auROC) curve (Green and Swets, 1966), which indicates the probability that an ideal observer could correctly classify the trial type to which a randomly selected response belongs. The auROC values, which range from zero to one, were transformed (discriminibility index = (auROC – 0.5) * 2) such that an index approaching one indicates that the dopamine response to a HM trial can be reliably discriminated as greater than the response to a LL trial, an index of zero indicates that the responses from each trial type cannot be discriminated, and an index approaching -1 would indicate greater dopamine release to the LL than to the HM option. This discriminability index timecourse confirmed that the cue-evoked dopamine response to the HM option was significantly greater than the response to the LL option, an effect we observed regardless of the side to which each option was assigned in the operant chamber (i.e., across the counterbalanced pairs of recorded sessions within this group of rats; Figure 2.3*A*). The greater dopamine response to the preferred HM option is consistent with numerous previous observations that options with greater subjective value are associated with greater dopamine-

reported cached values than are less-preferred options (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Roesch et al., 2007; Kobayashi and Schultz, 2008; Bromberg-Martin and Hikosaka, 2009; Day et al., 2010; Gan et al., 2010; Nasrallah et al., 2011; Sugam et al., 2012; Pasquereau and Turner, 2013; Lak et al., 2014; Stauffer et al., 2014). However, in these Moderate Cost sessions, animals' preferences are predominantly driven by the HM option's greater reward value, a dimension that is robustly incorporated into dopamine-associated cached values (Tobler et al., 2005; Roesch et al., 2007; Bromberg-Martin and Hikosaka, 2009; Gan et al., 2010; Nasrallah et al., 2011; Pasquereau and Turner, 2013; Stauffer et al., 2014). Thus, this condition demonstrated that these signals encode value-related information but did not allow us to determine whether dopamine-reported cached values still correlate with subjective value when preferences are driven by a weakly encoded economic dimension.

The critical test of this hypothesis was provided by the High Cost sessions in which animals reached the behavioral criterion for preferring the LL option, as the option yielding a larger reward was rendered the non-preferred option due to its high effort requirement, an attribute to which cue-evoked dopamine was relatively insensitive in previous studies (Ravel and Richmond, 2006; Gan et al., 2010; Wanat et al., 2010; Pasquereau and Turner, 2013). If the cached value signaled by cue-evoked dopamine reliably reflects subjective value, we would expect greater dopamine release to the preferred LL option. On the other hand, if the dopamine-reported cached value is more sensitive to expected reward value than to anticipated effort, we would expect a greater response to the HH option yielding a larger reward, despite the animals' subjective preferences for the LL alternative. In these High Cost sessions, cue-evoked dopamine release in the nucleus accumbens core was discriminable between HH and LL Forced trials, but only after

the peak response (Figures 2.2*E* and 2.3*B*). Remarkably, the greater of these responses was for the HH option, even though the LL option was significantly preferred. Therefore, in these High Cost sessions, the relative ordering of the cached values reported by cue-evoked dopamine release was inconsistent with the subjective value of each option. In other High Cost sessions where rats failed to reach criterion for the LL option but instead preferred the HH option (Figure 2.2*F*; $t_7 = 5.380$, $P = 0.001$), dopamine release was again greater for the HH option (Figure 2.2*G*). Thus, across all session types there was a greater dopamine-associated cached value for the high-value option (HM or HH) than for the LL option regardless of whether the cost to obtain this high-value option was only moderately higher or was much higher and regardless of whether or not it was preferred over the LL option.

The cue-evoked mesolimbic dopamine response was consistently more sustained for the high- than low-value option. However, during this sustained response, the remaining cost to obtain the reward becomes incrementally smaller with each lever press, and so this encoding pattern may reflect the dynamically increasing subjective value. To rule out this possibility, we conducted another set of experiments that allowed us to assess the cached values reported via cue-evoked dopamine release in a 5-s period before the lever was available (Figure 2.4*A*-*C*; significant post-criterion behavioral preferences in Moderate Cost sessions: $t_7 = 12.98$, $P = 3.74 \times 10^{-6}$; HH-preferred High Cost sessions: $t_9 = 5.267$, $P = 0.0005$; and LL-preferred High Cost sessions: $t_{11} = 4.319$, $P = 0.0012$). In this interval between cue and lever presentation, the pattern of dopamine responses for each of the sessions was comparable to the previous results, in that mesolimbic dopamine release was greater for the option yielding the high-value reward in all session types regardless of whether it was preferred or not (Figure 2.4*D*-*F*). Importantly, there was significant

discriminability between the high- and low-value options prior to lever presentation, demonstrating that the greater sustained response observed in high-value trials was not due to its increasing subjective value as the remaining response cost was reduced with each lever press.

To further examine the relationship between these behavioral and neurochemical data, we pooled the data for each session across all of the conditions in all of the rats. We plotted the dopamine discriminability index (the ability to discern the dopamine signal over the 5 s following cue onset as being greater for the larger- than the smaller-reward option) as a function of behavioral preferences in each session, using a choice index from post-criterion behavior: choice index = (p(H) – 0.5) * 2, where p(H) is the proportion of choices for the high-value option (HM or HH). Thus a choice index of 1 corresponds to 100% choice for the high-value option, -1 indicates 100% choice for the LL option, and 0 indicates indifference between the two options. This analysis allowed us to test whether the dopamine-reported cached values reflect subjective value. According to this prevailing hypothesis, the data should exclusively occupy the upper-right and lower-left quadrants of the graph. That is, the dopamine discriminability index should be positive when rats preferred the high-value option (positive choice index) and negative when rats preferred the LL option (negative choice index). Indeed the majority of points with a positive choice index were in the upper-right quadrant (43 of 44 sessions), which is significantly higher than expected by chance ($P = 5.16 \times 10^{-12}$, binomial test). However, when examining sessions with a negative choice index, the data diverged from this model, as the majority of these sessions did not have a negative dopamine discriminability index. In fact, a significant majority of the sessions with a negative choice index had a positive dopamine discriminability index (19 out of 23 sessions, $P = 0.0026$, binomial test; Figure 2.5*A*), favoring an alternative model where the

dopamine discriminability index is positive regardless of the animal's preference (62 out of 67 sessions, $P = 1.42 \times 10^{-13}$, binomial test; Figure 2.5*B*).

We next carried out a more detailed analysis of a utility model relating the dopamine-reported cached values to choice and compared this to a model where utility has no influence. We constructed the utility model as a regression line ($y = \beta_1 * x + \beta_0$) constrained through the origin ($\beta_0 = 0$; Figure 2.5*C*) and the alternative model as a constant with no slope ($\beta_1 = 0$; Figure 2.5*D*). Comparing the goodness-of-fit using the second-order bias-corrected Akaike Information Criterion (AICc) (Akaike, 1974; Burnham and Anderson, 2002) based on the residual sum of squares, the utility model provided an inferior fit to the data than did the alternative model which explicitly did not account for utility (AICc = -97.30 vs -154.27, respectively; weight of evidence favoring the origin-constrained slope = 4.26×10^{-13}, vs. > 0.999 favoring the constant with no slope). Although the positive bias in the dopamine discriminability index, independent of preference, observed in the alternative model was evident in all session types (Figure 2.5*E*; one-sample t-tests vs. zero: Moderate Cost sessions, $t_{19} = 11.90$, $P = 2.98 \times 10^{-10}$; HH-preferred High Cost sessions, $t_{21} = 7.78$, $P = 1.28 \times 10^{-7}$; LL-preferred High Cost sessions, $t_{24} = 4.67$, $P = 9.61 \times 10^{-5}$), its magnitude differed between these conditions (one-way ANOVA: $F_{2,64} = 4.28$, $P = 0.018$; Bonferroni-corrected post-hoc tests: Moderate Cost vs. LL-preferred High Cost sessions, $P = 0.015$; LL-preferred vs. HH-preferred High Cost sessions, $P = 0.312$; Moderate Cost vs. HH-preferred High Cost sessions, $P = 0.638$). Indeed, a standard linear regression model including both a slope and intercept term as free parameters provided an improved fit (Figure 2.5*F;* AICc = -163.45) over either the origin-constrained slope (Figure 2.5*C*), the constant line without slope (Figure 2.5*D*), or discontinuous lines (Figure 2.6 and Table 2.1). In this unconstrained model, the

slope of the linear term significantly differed from zero ($\beta_1 = 0.199 \pm 0.057$, $t = 3.468$, $P = 9.34 \times 10^{-4}$) and explained a small proportion of the variance in the dopamine discriminability index ($r^2 = 0.156$). Nonetheless, the y-intercept of this regression model also was significantly greater than zero ($\beta_0 = 0.403 \pm 0.038$, $t = 10.737$, $P = 4.89 \times 10^{-16}$), suggesting that behaviorally indifferent rats show greater cue-evoked dopamine release to a high- versus low-value option despite their lack of preference. Moreover, this regression line remained positive for all possible choice indices, meaning that dopamine release is greater for the high-value option regardless of preference. Importantly, the type of model that provided the best fit was not changed if we used the data from only the cohort of animals with the 5-s cue-to-lever delay (Figure 2.7 and Table 2.2), if the counterbalanced session pairs were treated as independent data points (Figure 2.8 and Table 2.3), or if we analyzed the peak dopamine release rather than the auROC-based discriminability index (Figure 2.9 and Table 2.4). Collectively, these data indicate that, although the dopamine-reported cached values showed a modest correlation with utility in these experimental paradigms, this relationship is not sufficient to confer them as a reliable instrument for determining choices.

Discussion

A preponderance of evidence supports the notion that phasic dopamine transmission functions as a neural instantiation of the temporal-difference prediction errors that drive reinforcement learning (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Waelti et al., 2001; Bayer and Glimcher, 2005; Roesch et al., 2007; Glimcher, 2011; Cohen et al., 2012; Schultz,

2013; Steinberg et al., 2013; Hart et al., 2014; Stopper et al., 2014). Accordingly, changes in dopamine transmission are evoked whenever there is an unexpected reward-related event, both when reward delivery differs from one's expectations and when reward-predictive cues drive changes in expectation of available reward. The latter exemplifies how the dopamine response to unexpected cue presentation provides a readout of the cached value assigned to that cue through temporal-difference learning. These cached values are theorized to be used to determine action selection, where the preferred action is the one associated with the cue with the greatest cached value (Sutton and Barto, 1998; Daw and Doya, 2006; Rangel et al., 2008; Kable and Glimcher, 2009; Lee et al., 2012). To subserve this role in decision making, the cached values need to incorporate any and all economic attributes insofar as those attributes influence subjective preferences; that is, by definition, the cached values must reliably reflect the ordinal utility of the available actions as revealed by animals' behavioral preferences. Consistent with this premise, there have been numerous reports in which dopamine-associated cached values incorporate many economic attributes affecting animals' behavioral preferences such as objective expected value (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Roesch et al., 2007; Gan et al., 2010; Pasquereau and Turner, 2013) and some subjective attributes including risk preference (Nasrallah et al., 2011; Sugam et al., 2012; Lak et al., 2014; Stauffer et al., 2014), temporal discounting (Roesch et al., 2007; Kobayashi and Schultz, 2008; Day et al., 2010), flavor or reward-type preference (Lak et al., 2014), perceptual uncertainty (Nomoto et al., 2010; de Lafuente and Romo, 2011), and even preference for advanced information (Bromberg-Martin and Hikosaka, 2009). Even though this system is based upon a cached-value ("model-free") architecture, there have been suggestions in the literature that it has access to "model-based" information derived from inferential online computation (Bromberg-Martin et al., 2010b), further

supporting the notion that the dopamine-associated cached values are a common currency for economic decision making (Schultz, 2013). But no matter how many positive correlations between dopamine-associated cached values and subjective preferences are observed, the existence of counterexamples where this relationship is reversed is sufficient to demonstrate that the fundamental claim of decision-making theories, that cached values are all that is required to determine action selection, simply cannot hold as a general principle. Accordingly, the current work identifies circumstances where there was a significant inversion between the options' ordinal utility and the rank ordering of their dopamine-reported cached values. This breakdown in the relationship between cached values and subjective value arose in situations where the animals' preferences were predominantly guided by effortful response cost, an economic dimension which previous voltammetry (Gan et al., 2010; Wanat et al., 2010) and electrophysiology (Ravel and Richmond, 2006; Pasquereau and Turner, 2013) studies have found to be weakly incorporated into the dopamine-associated cached values. Therefore, despite robust evidence that the dopamine-associated cached values do incorporate several subjective attributes such as risk preference (Nasrallah et al., 2011; Sugam et al., 2012; Lak et al., 2014; Stauffer et al., 2014) and temporal discounting (Roesch et al., 2007; Kobayashi and Schultz, 2008; Day et al., 2010), the current results by necessity imply that these cached values alone are insufficient for determining economic choices.

We have demonstrated that dopamine-associated cached values are positively correlated with behavioral preferences in some circumstances – those where the benefits overshadowed the costs – but are diametrically opposed to preference order in others – where the costs predominated in guiding animals' choices. Therefore, valuations used in decision making cannot be based on this

simple cached-value-based system alone, as they need to account for costs that influence action selection. Theorists have proposed the competition for the control of behavioral resources by co-existing valuation systems (Sutton and Barto, 1998; Daw et al., 2005; Daw and Doya, 2006; Rangel et al., 2008; Kable and Glimcher, 2009; Lee et al., 2012). Based on this competition framework, one could speculate that choices were determined by an alternative valuation system under circumstances when costs outweigh benefits (i.e., when the dopamine-associated cached values did not correlate with preference), and that the cached values were used as the determinant of choice only under circumstances when benefits overshadow costs. However, this marginalized use of the cached-value system would be quite limited because real-world choices are often strongly influenced by aversive or energetic costs. Moreover, even when benefits do overshadow costs, there are still gradations in preferences for incremental changes in response cost (Ghods-Sharifi et al., 2009), and so even these preferences cannot be based on a valuation system that is relatively insensitive to costs. Therefore, additional information on costs is required to perform these decision-making computations. Representations of costs have indeed been observed in areas such as the anterior cingulate cortex (Walton et al., 2003; Rudebeck et al., 2006; Hillman and Bilkey, 2010; Kennerley et al., 2011; Amemori and Graybiel, 2012), insular cortex (Prévost et al., 2010; Palminteri et al., 2012), and basolateral amygdala (Ghods-Sharifi et al., 2009; McHugh et al., 2014), all of which provide glutamatergic inputs that converge on striatal projection neurons. A more integrative valuation system, therefore, could arise from the downstream combination of benefit information from the dopamine-associated cached values and cost information from other neural sources. However, this concept of an incomplete valuation system requiring additional information is not accommodated in current theories (Sutton and Barto, 1998; Daw et al., 2005; Daw and Doya, 2006; Rangel et al., 2008; Kable and

Glimcher, 2009; Lee et al., 2012) whether they describe the dopamine-associated valuation system alone or used in parallel with alternative systems. Alternatively, it remains possible that dopamine-associated cached values do not contribute to the selection process at all, but rather play a more nuanced role in decision making that pertains to the performance or execution of the selected action. This scenario would place the burden of action selection on other reward-related structures. Indeed, representations of subjective value have been observed in multiple cortical regions (Kim et al., 2008; Louie and Glimcher, 2010; Kennerley et al., 2011; Padoa-Schioppa, 2011). Perturbation of the cached values (Tai et al., 2012; Steinberg et al., 2013; Stopper et al., 2014) to test the extent of their contribution to action selection could discern between these possibilities in future experiments. Under any of these scenarios, though, the current findings demonstrate that the dopamine-associated cached values alone are not sufficient to serve as the basis of simple economic choices.

In the current work, we used cue-evoked mesolimbic dopamine transmission as a means to examine cached values assigned to reward options, and identified circumstances where the option that animals preferentially selected did not have the greatest cached value. This situation arose in the cost-benefit decisions of the present experiment when the differences in response costs overshadowed the differences in benefits. These findings demonstrate a direct violation of the fundamental principle that these cached values reflect animals' subjective preferences and are sufficient for determining choices. Therefore, we conclude that dopamine-associated cached values cannot be used as the sole determinant of cost-benefit decision making.

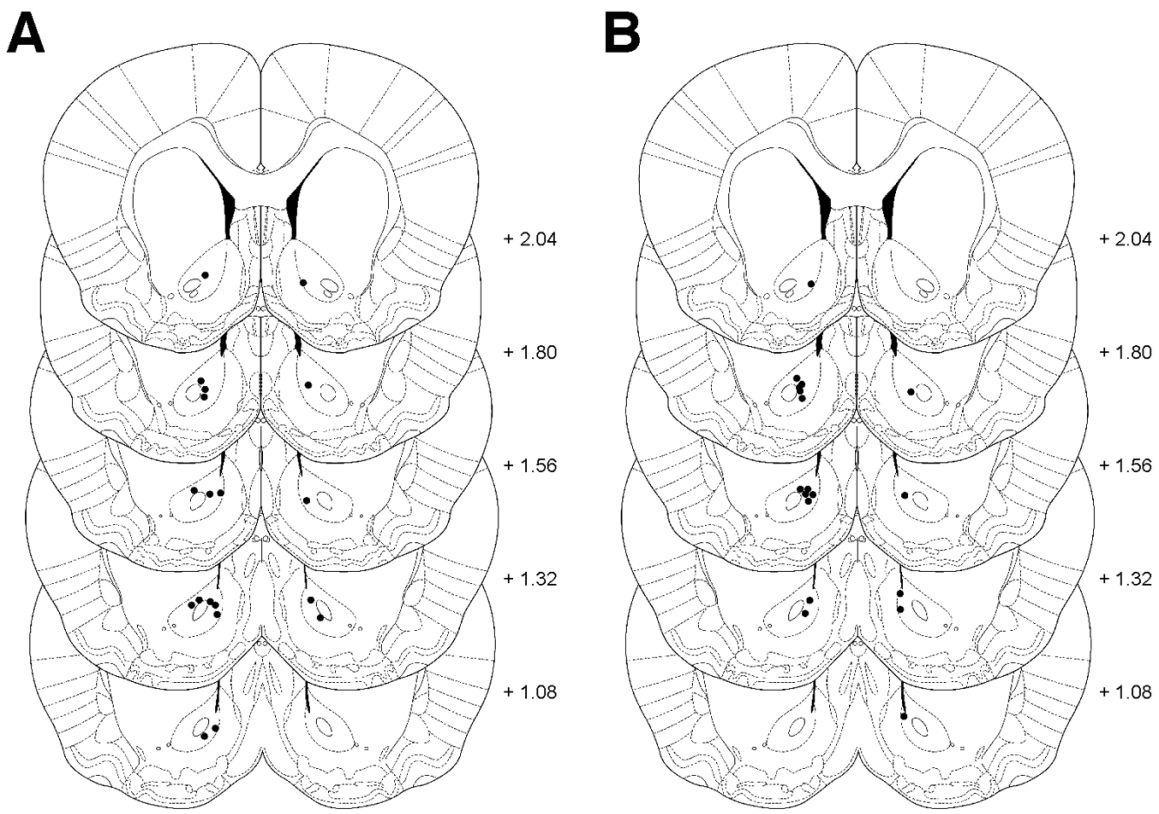

Figure 2.1. Recording locations in the nucleus accumbens core.

(*A*) The first and (*B*) the second cohorts of rats. The numbers next to each section indicate distance in mm anterior to bregma. Adapted from the atlas of Paxinos and Watson (2005).

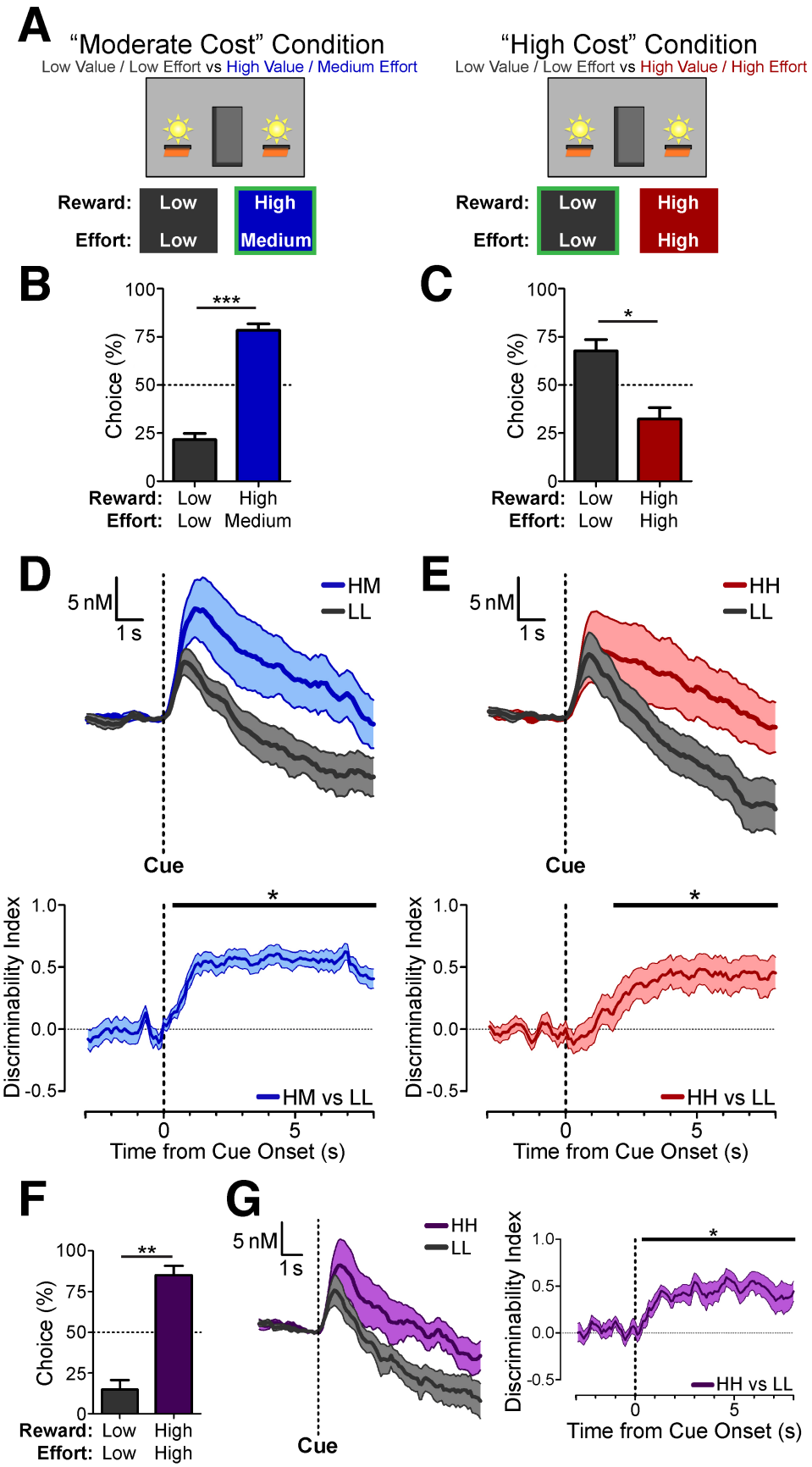
A
"Moderate Cost" Condition
Low Value / Low Effort vs High Value / Medium Effort
Reward: Low High
Effort: Low Medium
"High Cost" Condition
Low Value / Low Effort vs High Value / High Effort
Reward: Low High
Effort: Low High
B
Choice (%)
***
Reward: Low High
Effort: Low Medium
C
Choice (%)
*
Reward: Low High
Effort: Low High
D
5 nM
1 s
HM
LL
Cue
Discriminability Index
HM vs LL
Time from Cue Onset (s)
E
5 nM
1 s
HH
LL
Cue
Discriminability Index
HH vs LL
Time from Cue Onset (s)
F
Choice (%)
**
Reward: Low High
Effort: Low High
G
5 nM
1 s
HH
LL
Cue
Discriminability Index
HH vs LL
Time from Cue Onset (s)

Figure 2.2. Task design, behavioral performance, and voltammetry results from forced trials with simultaneous cue and lever onset (first cohort).

(*A*) Reward and effort contingencies for Moderate Cost and High Cost sessions (green box signifies preferred option). (*B* and *C*) Mean (+ SEM) post-criterion percent choice for (*B*) Moderate Cost sessions (*** $P = 0.0001$, $n = 9$ rats) and (*C*) High Cost sessions (* $P = 0.0265$, $n = 7$ rats). (*D* and *E*) Mean (± SEM) cue-evoked dopamine release (*Upper*) and discriminability index time series (*Lower*) in Moderate Cost sessions (*D,* $n = 11$ recording sites) and in LL-preferred High Cost sessions (*E*, $n = 10$ recording sites). (*F*) Mean (+ SEM) post-criterion percent choice (** $P = 0.001$, $n = 8$ rats), and (*G*) mean (± SEM) cue-evoked dopamine release (*Left*) and discriminability index time series (*Right*) in HH-preferred High Cost sessions ($n = 11$ recording sites). For each discriminability index time course in *D*, *E*, and *G*, horizontal bar indicates time points of significant discriminability (* $P < 0.05$, permutation tests).

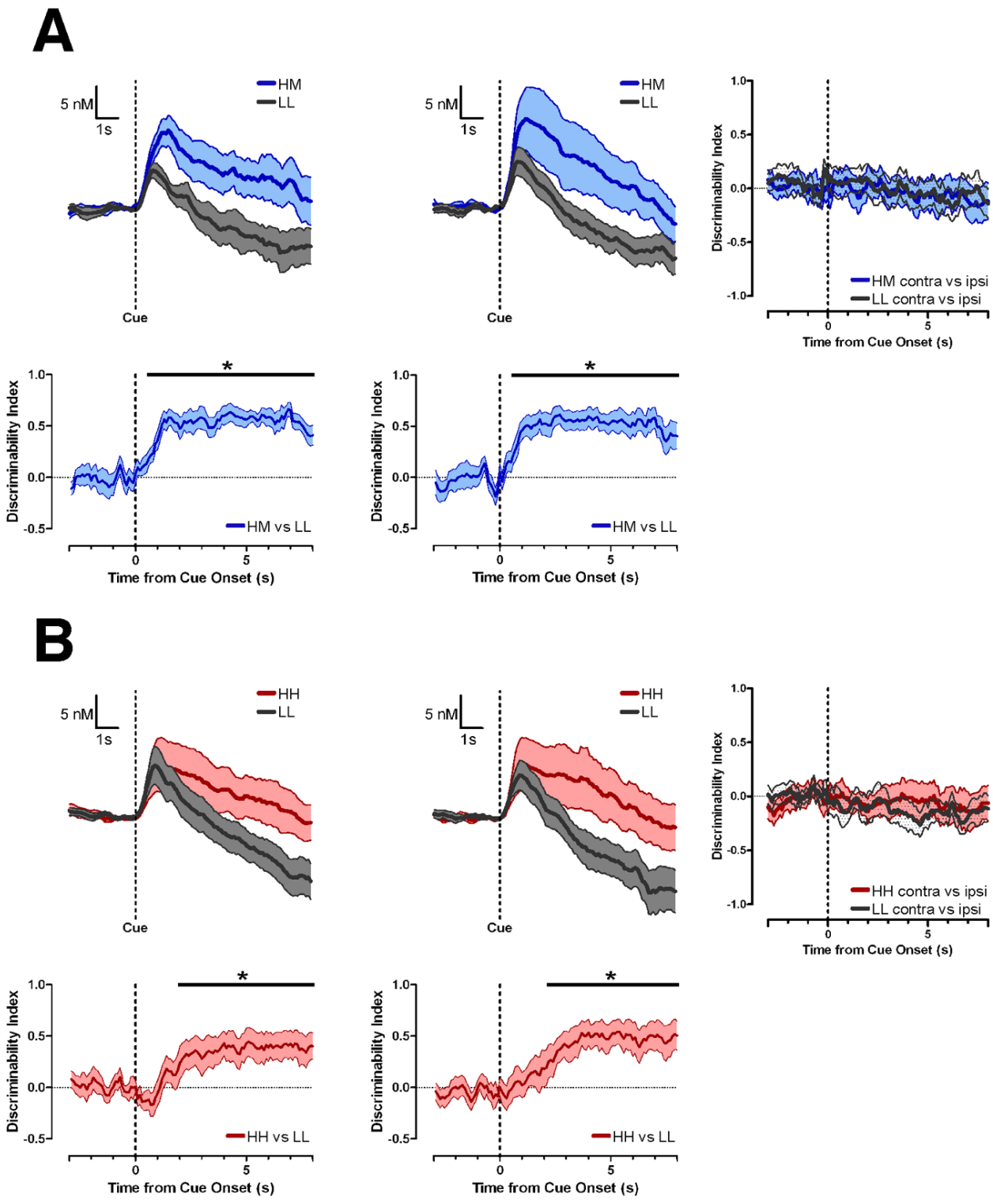
A
5 nM
1s
HM
LL
Cue
Discriminability Index
HM vs LL
HM contra vs ipsi
LL contra vs ipsi
Time from Cue Onset (s)
B
HH
HH vs LL
HH contra vs ipsi

Figure 2.3. Lack of direction-selective encoding by mesolimbic dopamine.

(*A*, *Upper*) Mean (± SEM) cue-evoked dopamine release during Forced trials from the side-counterbalanced pairs of Moderate Cost sessions recorded in rats in the first cohort, separated by lever side assignment. (*Left*) Sessions where the HM option (blue) was assigned to the lever ipsilateral to the hemisphere of the carbon-fiber microelectrode and the LL option (gray) was contralateral. (*Middle*) Sessions where the HM option (blue) was assigned to the lever contralateral to the hemisphere of the electrode and the LL option (gray) was ipsilateral. (*Right*) Mean (± SEM) discriminability index time series comparing Forced trials from the side-counterbalanced pairs. Neither the HM contralateral vs. HM ipsilateral comparison (blue) nor the LL contralateral vs. LL ipsilateral comparison (gray) ever reached significance. (*Lower*) Mean (± SEM) discriminability index time series comparing HM vs. LL Forced trials within each session, with lever assignments defined as above. Horizontal bars indicate time points of significant discriminability (* $P < 0.05$, permutation tests). (*B*, *Upper*) Mean (± SEM) cue-evoked dopamine release during Forced trials from the side-counterbalanced pairs of High Cost sessions recorded in rats in the first cohort reaching behavioral criterion for preferring the LL option, separated by lever side assignment. (*Left*) Sessions where the HH option (red) was assigned to the lever ipsilateral to the hemisphere of the carbon-fiber microelectrode and the LL option (gray) was contralateral. (*Middle*) Sessions where the HH option (red) was assigned to the lever contralateral to the hemisphere of the electrode and the LL option (gray) was ipsilateral. (*Right*) Mean (± SEM) discriminability index time series comparing Forced trials from the side-counterbalanced pairs. Neither the HH contralateral vs. HH ipsilateral comparison (red) nor the LL contralateral vs. LL ipsilateral comparison (gray) ever reached significance. (*Lower*) Mean (± SEM) discriminability index time series comparing HH vs. LL Forced trials within each session, with lever assignments defined as above. Horizontal bars indicate time points of significant discriminability (* $P < 0.05$, permutation tests).

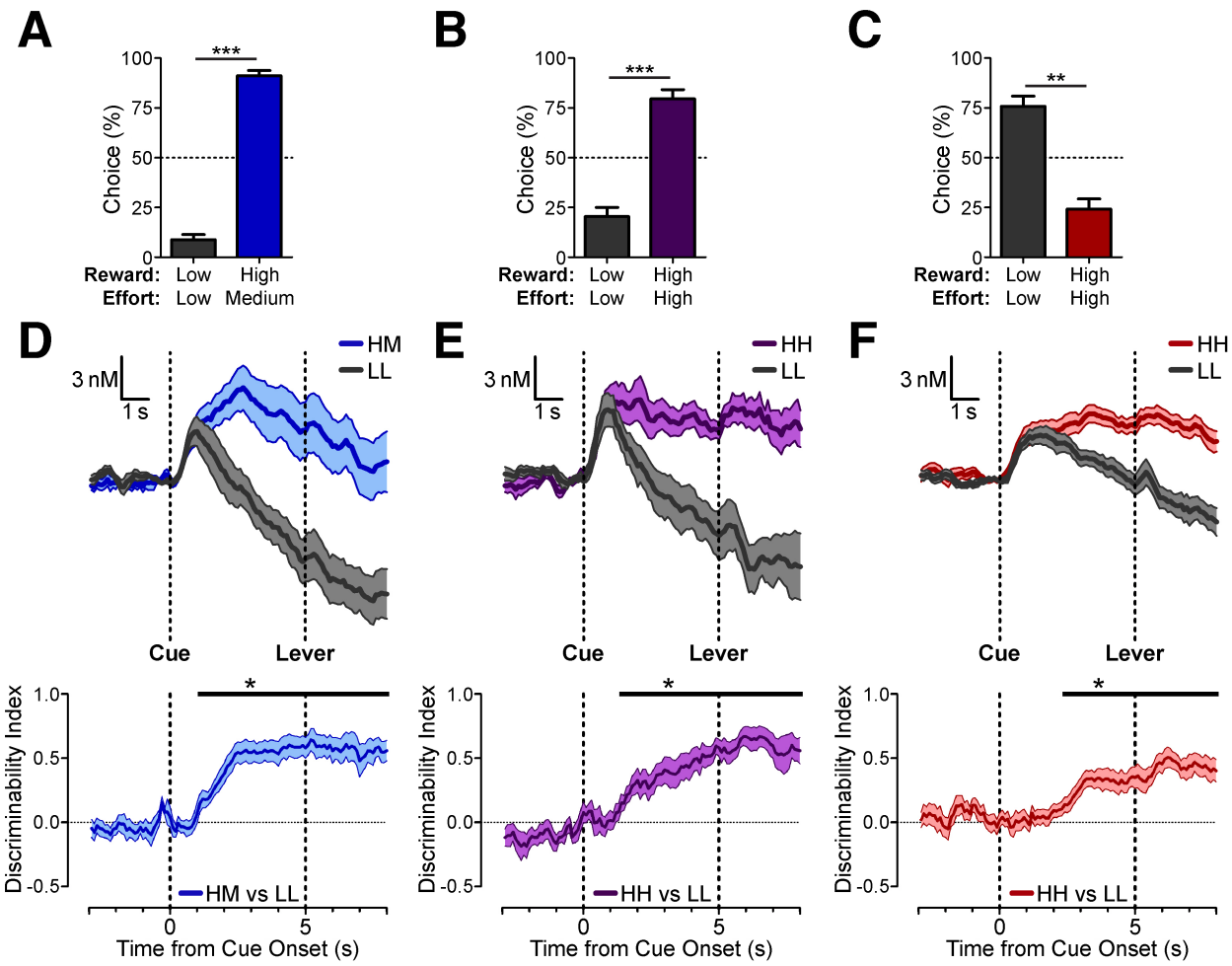

Figure 2.4. Behavioral performance and voltammetry results from forced trials with a 5 s cue-to-lever delay (second cohort).

(*A* to *C*) Mean (+ SEM) post-criterion percent choice for (*A*) Moderate Cost sessions (*** $P = 3.74 \times 10^{-6}$, $n = 8$ rats), (*B*) HH-preferred High Cost sessions (*** $P = 0.0005$, $n = 10$ rats), and (*C*) LL-preferred High Cost sessions (** $P = 0.0012$, $n = 12$ rats). (*D* to *F*) Mean (± SEM) cue-evoked dopamine release (*Upper*) and discriminability index time series (*Lower*) in Moderate Cost sessions (*D*, $n = 9$ recording sites), in HH-preferred High Cost sessions (*E*, $n = 11$ recording sites), and in LL-preferred High Cost sessions (*F*, $n = 15$ recording sites). For each discriminability index time course in *D* to *F*, horizontal bar indicates time points of significant discriminability (* $P < 0.05$, permutation tests).

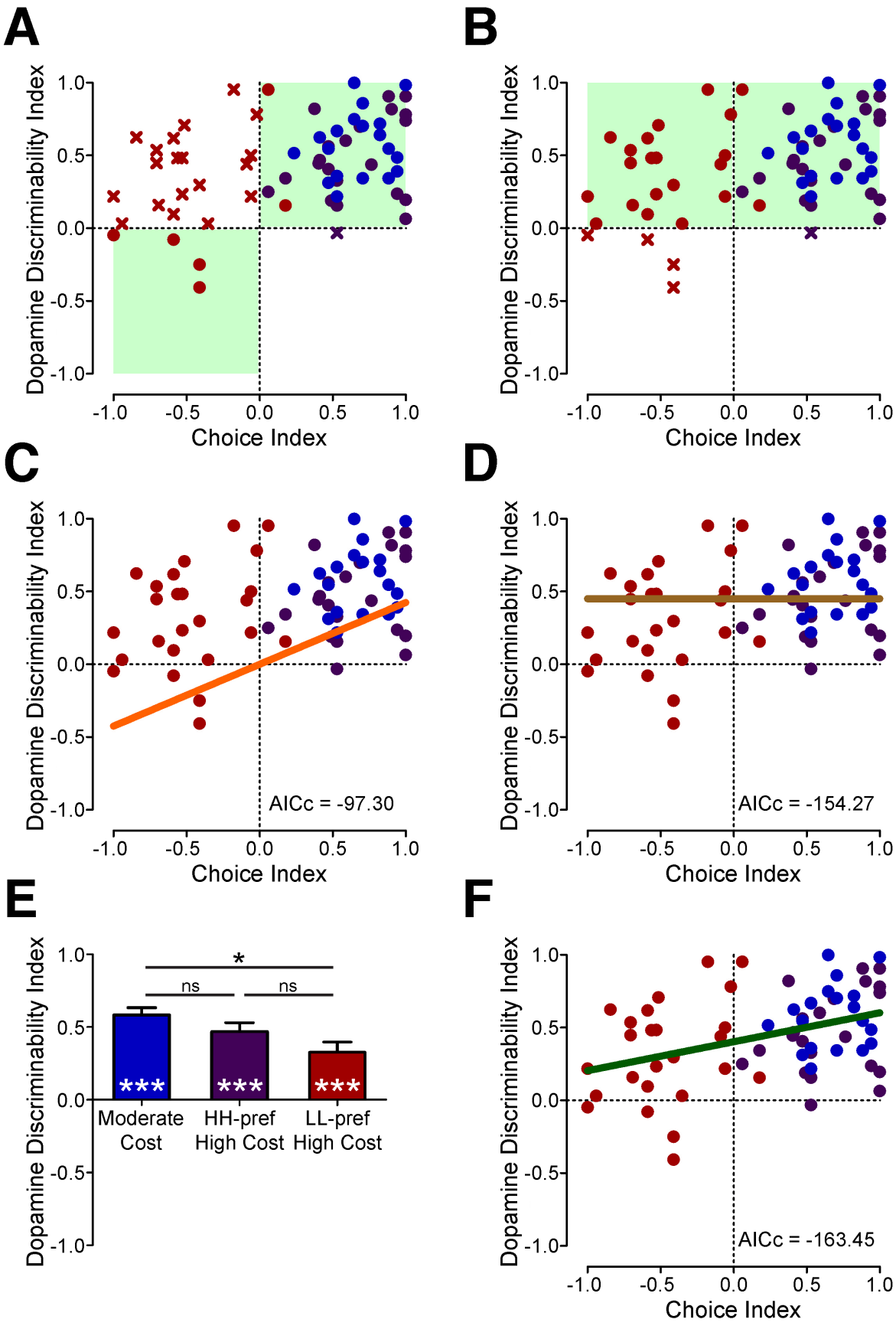
A
B
C
D
E
F
Dopamine Discriminability Index
Choice Index
1.0
0.5
0.0
-0.5
-1.0
AICc = -97.30
AICc = -154.27
AICc = -163.45
*
ns
ns
***
Moderate Cost
HH-pref High Cost
LL-pref High Cost

Figure 2.5. Models testing the relationship between dopamine-associated cached values and subjective preferences.

Throughout all panels, blue points are Moderate Cost sessions ($n = 20$), red points are LL-preferred High Cost sessions ($n = 25$), and purple points are HH-preferred High Cost sessions ($n = 22$). X's designate sessions violating the categorical models' predictions (*A* and *B*). (*A*) Expected utility categorical model predicts that all data should fall within the upper-right and lower-left quadrants (green). (*B*) Expected benefits categorical model predicts that all data should fall within the upper quadrants (green). (*C*) Expected utility regression model: linear regression constrained through origin ($y = \beta_1 * x$; $\beta_0 = 0$). (*D*) Expected benefits regression model: constant line, no slope ($y = \beta_0$; $\beta_1 = 0$). (*E*) The average dopamine discriminability index was significantly greater than zero for each session type (*** $P < 0.001$) but was lower in the LL-preferred High Cost sessions than in the Moderate Cost sessions (* $P = 0.015$). (*F*) Standard linear regression model ($y = \beta_1 * x + \beta_0$). Both the slope and intercept significantly differ from zero ($P < 0.001$).

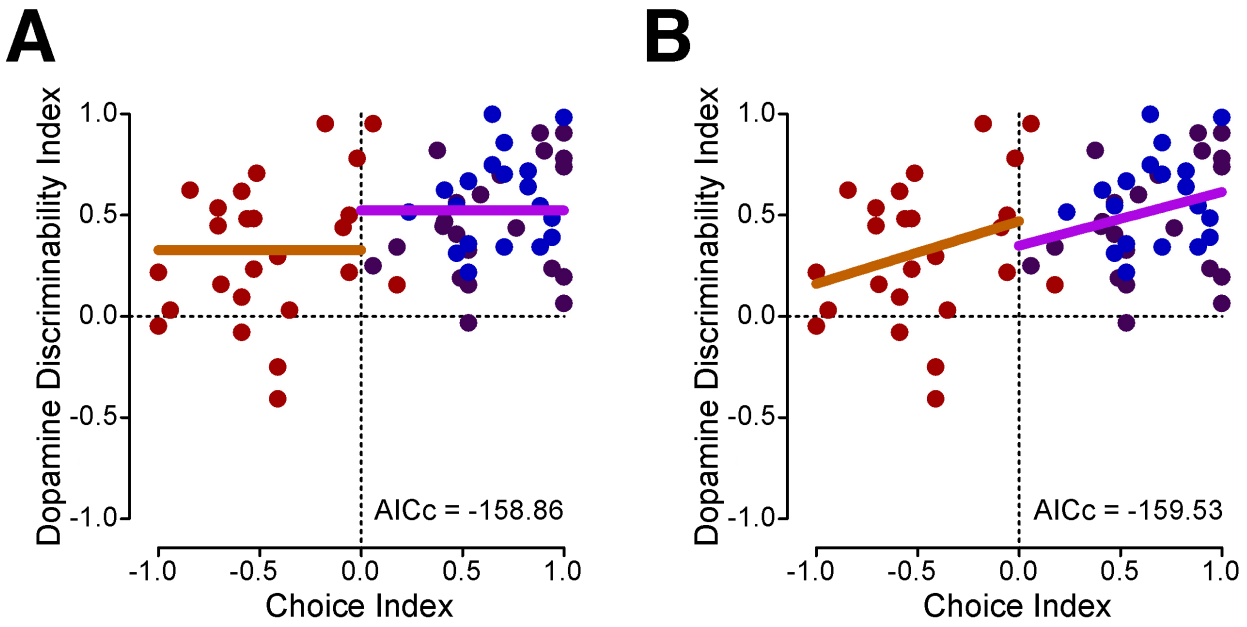

Figure 2.6. Additional models testing the relationship between dopamine-associated cached values and subjective preferences.

(*A*) Modeling the discriminability index with two constants, depending on whether the low- or high-value option was preferred (AICc = -158.86). (*B*) Separate linear regressions for when the low- vs. high-value option was preferred (AICc = -159.53). Each provided a better fit than either the origin-constrained utility model (AICc = -97.30, Figure 2.5*C*) or the single-constant without slope (AICc = -154.27, Figure 2.5*D*), but was not better than the standard linear regression model (AICc = -163.45, Figure 2.5*F*). As in Figure 2.5, blue points represent Moderate Cost sessions, red points are LL-preferred High Cost sessions, and purple points are HH-preferred High Cost sessions.

Table 2.1. AICc and weights of evidence for each model from Figures 2.5 and 2.6.

Model	Free Parameters	AICc	Weight for Model
Origin-constrained slope	2	-97.30	3.45×10^{-15}
Constant with no slope	2	-154.27	0.0081
Standard linear regression	**3**	**-163.45**	**0.7991**
Two constants	3	-158.86	0.0802
Two linear regressions	5	-159.53	0.1125

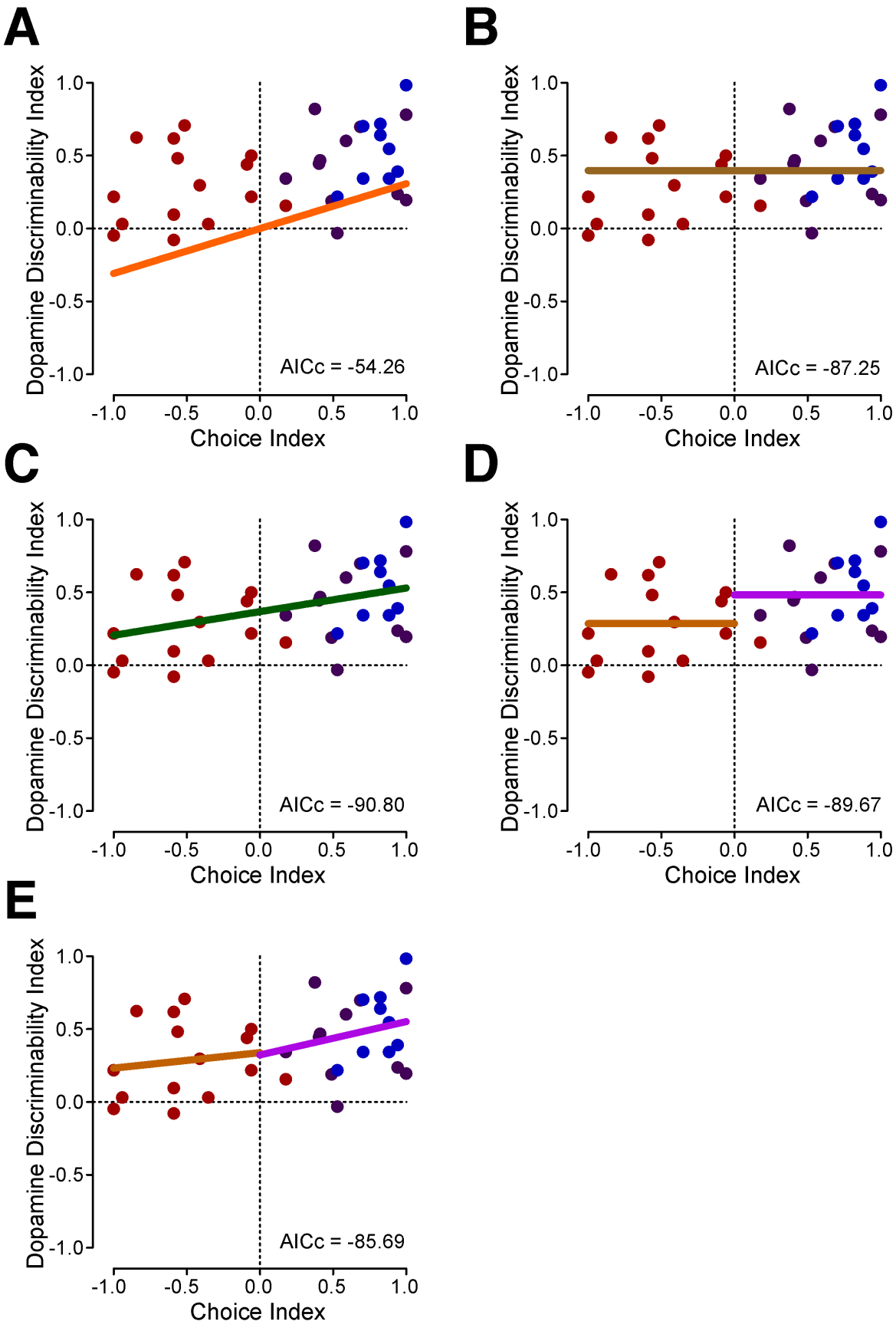
A
B
C
D
E
Dopamine Discriminability Index
Choice Index
1.0
0.5
0.0
-0.5
-1.0
AICc = -54.26
AICc = -87.25
AICc = -90.80
AICc = -89.67
AICc = -85.69

Figure 2.7. Models from Figures 2.5 and 2.6, including only the data from the second cohort (5 s cue-to-lever delay).

(*A*) Expected utility regression model: constrained through origin (AICc = -54.26). (*B*) Constant line without slope (AICc = -87.25). (*C*) Standard linear regression (AICc = -90.80; $r^2 = 0.1564$). Both the slope and intercept significantly differ from zero ($\beta_1 = 0.162 \pm 0.066$, $t = 2.473$, $P = 0.019$; $\beta_0 = 0.368 \pm 0.045$, $t = 8.210$, $P = 1.76 \times 10^{-9}$). (*D*) Two constants, depending on whether the low- or high-value option was preferred (AICc = -89.67). (*E*) Separate linear regressions for when the low- vs. high-value option was preferred (AICc = -85.69). Throughout all panels, blue points are Moderate Cost sessions, red points are LL-preferred High Cost sessions, and purple points are HH-preferred High Cost sessions.

Table 2.2. AICc and weights of evidence for each model from Figure 2.7.

Model	Free Parameters	AICc	Weight for Model
Origin-constrained slope	2	-54.26	6.39×10^{-9}
Constant with no slope	2	-87.25	0.0932
Standard linear regression	**3**	**-90.80**	**0.5511**
Two constants	3	-89.67	0.3130
Two linear regressions	5	-85.69	0.0427

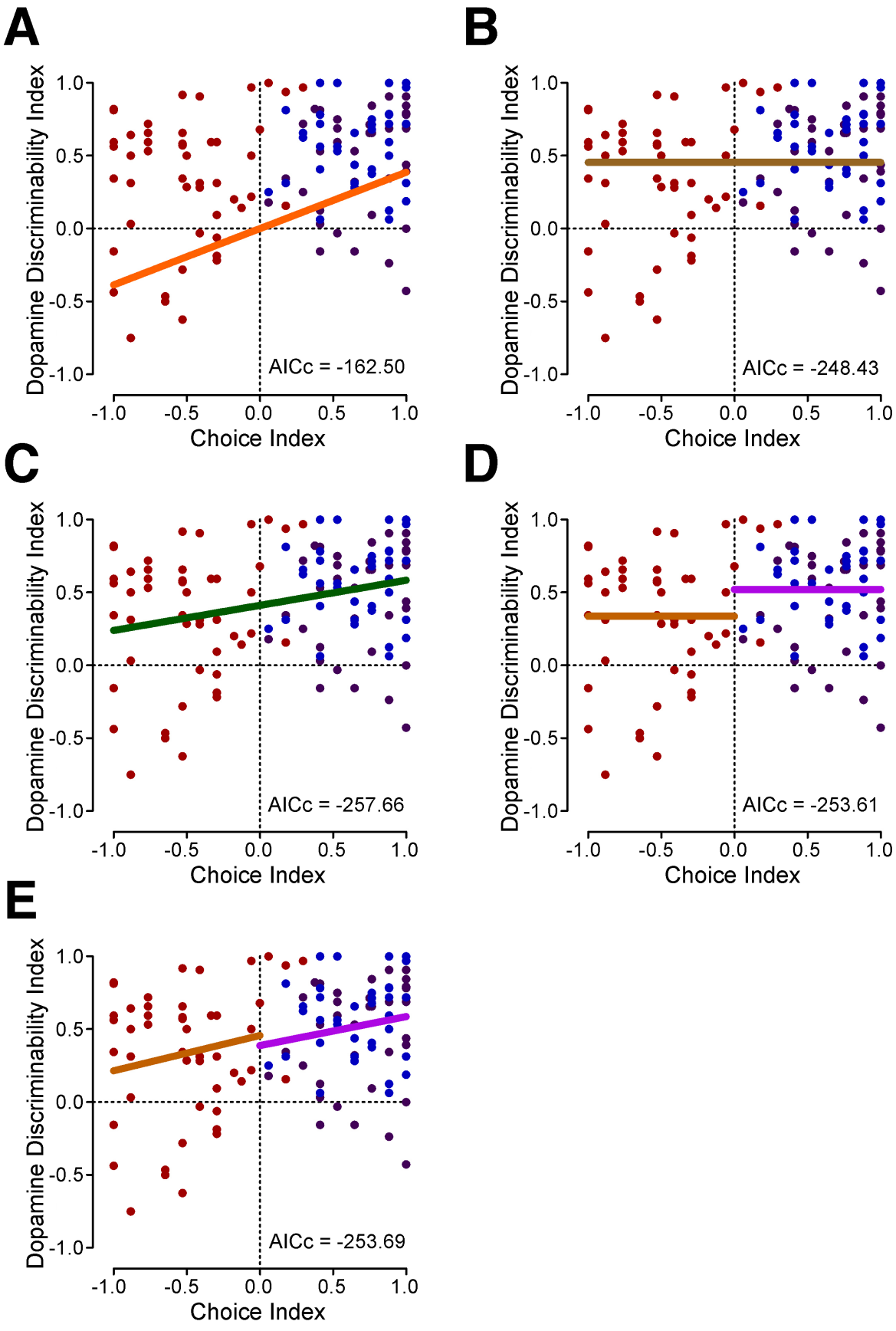
A
B
C
D
E
Dopamine Discriminability Index
Choice Index
1.0
0.5
0.0
-0.5
-1.0
AICc = -162.50
AICc = -248.43
AICc = -257.66
AICc = -253.61
AICc = -253.69

Figure 2.8. Models from Figures 2.5 and 2.6, splitting the pairs of counterbalanced sessions and treating each as an independent data point.

(*A*) Expected utility regression model: constrained through origin (AICc = -162.50). (*B*) Constant line without slope (AICc = -248.43). (*C*) Standard linear regression (AICc = -257.66; $r^2 = 0.0822$). Both the slope and intercept differ significantly from zero ($\beta_1 = 0.172 \pm 0.050$, $t = 3.413$, $P = 8.56 \times 10^{-4}$; $\beta_0 = 0.411 \pm 0.035$, $t = 11.900$, $P = 1.52 \times 10^{-22}$). (*D*) Two constants, depending on whether the low- or high-value option was preferred (AICc = -253.61). (*E*) Separate linear regressions for cases in which the low- vs. high-value option was preferred (AICc = -253.69). In all panels, blue points are moderate-cost sessions, red points are LL-preferred high-cost sessions, and purple points are HH-preferred high-cost sessions.

Table 2.3. AICc and weights of evidence for each model from Figure 2.8.

Model	Free Parameters	AICc	Weight for Model
Origin-constrained slope	2	-162.50	1.69×10^{-21}
Constant with no slope	2	-248.43	0.0077
Standard linear regression	**3**	**-257.66**	**0.7821**
Two constants	3	-253.61	0.1029
Two linear regressions	5	-253.69	0.1073

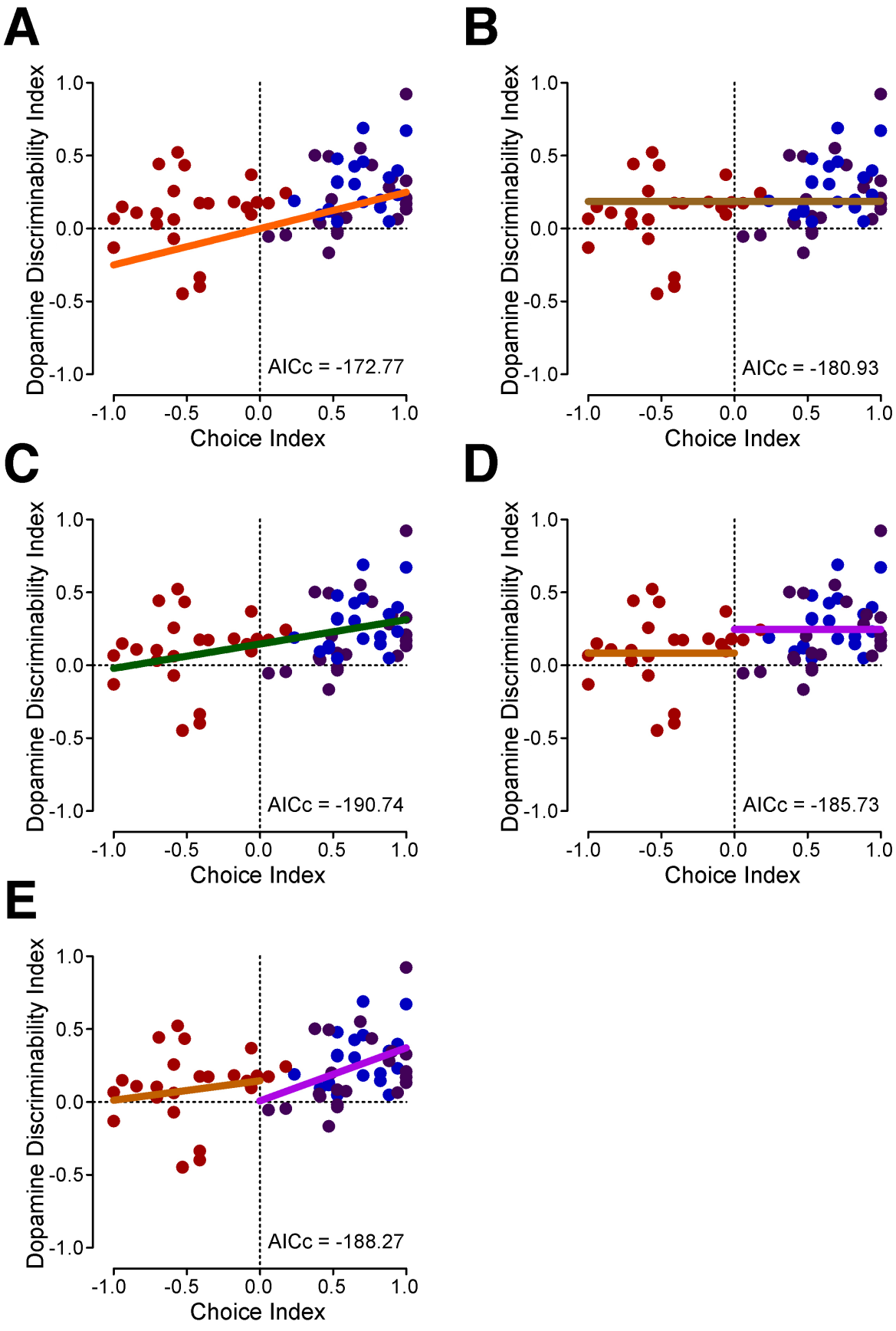
A
B
C
D
E
Dopamine Discriminability Index
Choice Index
1.0
0.5
0.0
-0.5
-1.0
AICc = -172.77
AICc = -180.93
AICc = -190.74
AICc = -185.73
AICc = -188.27

Figure 2.9. Models from Figure 2.5 and 2.6, using a peak dopamine index [(H-L)/(H+L)] instead of the auROC-based discriminability index.

(*A*) Expected utility regression model: constrained through origin (AICc = -172.77). (*B*) Constant line without slope (AICc = -180.93). (*C*) Standard linear regression (AICc = -190.74; $r^2 = 0.1639$). Both the slope and intercept differ significantly from zero ($\beta_1 = 0.167 \pm 0.047$, $t = 3.570$, $P = 6.77 \times 10^{-4}$; $\beta_0 = 0.146 \pm 0.031$, $t = 4.776$, $P = 1.06 \times 10^{-5}$). (*D*) Two constants, depending on whether the low- or high-value option was preferred (AICc = -185.73). (*E*) Separate linear regressions for cases in which the low- vs. high-value option was preferred (AICc = -188.27). In all panels, blue points are moderate-cost sessions, red points are LL-preferred high-cost sessions, and purple points are HH-preferred high-cost sessions.

Table 2.4. AICc and weights of evidence for each model from Figure 2.9.

Model	Free Parameters	AICc	Weight for Model
Origin-constrained slope	2	-172.77	9.11×10^{-5}
Constant with no slope	2	-180.93	0.0054
Standard linear regression	**3**	**-190.74**	**0.7244**
Two constants	3	-185.73	0.0593
Two linear regressions	5	-188.27	0.2108

Chapter 3

Cue-evoked mesolimbic dopamine signals the cached value of the chosen outcome during decisions between concurrently available options*

* This chapter is currently in preparation for submission with minor reformatting as an article the with the same title and Monica M. Arnold, Jerylin O. Gan, Mark E. Walton, and Paul E. M. Phillips as coauthors.

All authors contributed to experimental design. I collected the experimental data with assistance from M.M.A. and J.O.G., I carried out data analysis with assistance from M.M.A., and I prepared the manuscript with M.M.A. and P.E.M.P.

Abstract

By encoding the reward-prediction error term of temporal-difference reinforcement learning algorithms, phasic dopamine transmission is implicated in learning the value of reward-predictive states and actions. Upon presentation of a reward-predictive cue, dopaminergic prediction errors provide a readout of the associated cached value when a single action is available to obtain reward. Still controversial, however, is which cached value is reported by this dopamine signal upon presentation of cues denoting the opportunity to choose between concurrently available options, and therefore, which variant of temporal-difference reinforcement learning algorithms this neural system enacts. We used fast-scan cyclic voltammetry to measure dopamine release in the nucleus accumbens core of rats making choices between options yielding small and large rewards for different effortful response costs. During choices between concurrently available options, cue-evoked mesolimbic dopamine release reported the cached value of the chosen option, implying the instantiation of an “on-policy” learning algorithm.

Introduction

In reverse engineering the brain, a goal of neuroscience is to understand the computations performed by the central nervous system at the precise algorithmic level. The phasic activity of many midbrain dopamine-containing neurons encodes errors in reward prediction (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997), resembling the critical teaching signal in formal reinforcement learning models for updating the cached values assigned to actions yielding reward when taken from a given state (state-action values) (Sutton and Barto, 1998). Several variants of temporal-difference-reinforcement-learning (TDRL) algorithms have been described for updating state-action values using prediction errors (Sutton and Barto, 1998). "On-policy" TDRL algorithms such as SARSA (i.e., State-Action-Reward-State-Action) use prediction errors based on the cached value of the action actually selected (Rummery and Niranjan, 1994), whereas "off-policy" algorithms such as Q-learning use those based upon the greatest cached value available regardless of which action is selected from that particular state (Watkins, 1989). These classes of algorithms generate identical temporal-difference prediction errors in states in which only one action is available to obtain reward (e.g., in single-option forced trials) or when agents select the action with the greatest cached value during choices between concurrently available options. The prediction errors generated within these algorithms only diverge when agents select an action that is not associated with the greatest cached value. When this action is the non-preferred option, it is only chosen when agents distribute their choices across the available actions. However, the reason for choosing alternatives to a preferred option on some trials is quite controversial, with hypotheses including exploration bonuses attached to the non-preferred option (Dayan and Sejnowski, 1996; Kakade and Dayan, 2002), foraging policies such

as matching (Herrnstein, 1961), noisy neural representations of the values of the options (Gold and Shadlen, 2007), or simply mistakes on the part of the subject. If any of these putative processes are incorporated into the dopamine-encoded prediction error (Dayan and Sejnowski, 1996; Kakade and Dayan, 2002), differentiating the TDRL algorithms based on these anomalous choices will become problematic.

The few studies recording midbrain-dopamine-neuron activity when animals face choices between concurrently available options arrived at conflicting conclusions regarding which TDRL algorithm best describes these cells' phasic activity. Morris et al. (2006) found that the cue-evoked responses of dopamine neurons in the substantia nigra pars compacta (SNc) of macaques encoded the value of the chosen option regardless of the value of the forgone alternative, resembling a neural instantiation of a SARSA algorithm (Niv et al., 2006). In contrast, a subsequent study by Roesch et al. (2007), focusing on the ventral tegmental area (VTA) of rats, suggested that dopamine neurons signaled the greatest cached value available regardless of what the animal chooses on a given trial, favoring a Q-learning algorithm (Daw, 2007). This homogeneity of dopamine responses across choices for either option observed in the latter study (Roesch et al., 2007) has been corroborated by recordings of sub-second dopamine release in the nucleus accumbens core using fast-scan cyclic voltammetry (Day et al., 2010; Sugam et al., 2012). Notably, the relative magnitudes of dopamine transmission on forced trials and choices for the option with the greatest cached value are highly consistent across these studies despite differences in species, recording location, and behavioral task design. The discrepancies in the patterns of dopamine transmission only arose on trials where the option with the lower cached value was chosen, the critical trial-type for differentiating on- and off-policy algorithms.

Importantly, in all of these studies, the option with the lower cached value was the not the preferred option, and therefore, analysis of the critical choice trials is subject to the aforementioned concerns related to contamination of the prediction errors during anomalous choices.

Although in these studies, as well as many others, the preferred option had the greater dopamine-associated cached value (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Morris et al., 2006; Roesch et al., 2007; Kobayashi and Schultz, 2008; Bromberg-Martin and Hikosaka, 2009; Day et al., 2010; Nasrallah et al., 2011; Sugam et al., 2012; Lak et al., 2014), we recently demonstrated that these cached values can be dissociated from the ordinal utility of available options, providing circumstances in which the option with the smaller cached value was preferred (Hollon et al., 2014). This innovation, using our previously described mixed-contingency decision-making task in which choices involve cost-benefit tradeoffs (Hollon et al., 2014), permitted us to examine the computational algorithm enacted by mesolimbic dopamine transmission without relying upon data obtained from anomalous choices on trials where animals selected the non-preferred option.

Methods

Overview of Experiments. Choice behavior and single-option forced trial voltammetry results from experiments one and two have previously been described in Chapter 2 (Hollon et al., 2014); here we present novel analyses of these data sets to investigate distinct hypotheses pertaining to

mesolimbic dopamine transmission during choices between concurrently available options. Experiment three employs a new task variant run with a third group of rats to address alternative explanations of results from the first two experiments. All procedures were approved by the University of Washington Institutional Animal Care and Use Committee.

Subjects and surgery. A total of 52 male Sprague Dawley rats (Charles River Laboratories), 250-300 grams upon arrival, were used for this study. Fourteen, fourteen, and eleven rats contributed data to experiments one, two, and three, respectively; four additional rats were excluded due to electrode misplacement, three due to failure of electrodes to satisfy criteria for dopamine detection, and six due to post-surgical complications (e.g., head-cap loss). Rats were maintained on a 12-hour light/dark cycle, with all behavioral testing occurring during the light phase. Rats were pair-housed until surgery, after which they were housed individually. Rats were anesthetized with isoflurane for bilateral implantation of carbon-fiber microelectrodes (Clark et al., 2010) targeting the nucleus accumbens core (1.3 mm anterior, 1.3 mm lateral, 6.8-7.0 mm ventral to bregma, according to the atlas of Paxinos and Watson, 2005; Figures 2.1 and 3.3) and a Ag/AgCl reference electrode. After at least one week recovery post-surgery, rats were food-restricted to 90% their *ad libitum* body weight; for all subsequent behavioral procedures, each rat received a total of ~15 grams of food per day consisting of pellets earned as reward during behavioral sessions plus standard lab chow after these sessions. Water was available *ad libitum* in the animals' home cages.

Experiments one and two. Initial training and the mixed-contingency decision-making task have been described in the methods section of Chapter 2. Experiment one refers to the task

variant with simultaneous cue light onset and lever presentation, and experiment two refers to the variant with a 5-s cue-to-lever delay.

Experiment three: initial training. After magazine training and free-operant (continuous reinforcement FR-1) sessions for each lever as in the first two experiments, the third group of rats completed one session of discrete-trial FR-1 with a 4 ± 1 s ITI and simultaneous cue onset and lever presentation. For all subsequent sessions these rats were required make a head entry into the centrally located food magazine after the ITI period ended to initiate the next trial; the end of the ITI period was not signaled by any overt stimulus. These rats then completed one session of FR-1 with a 20 ± 5 s ITI. All subsequent sessions included the 10 s time limit to initiate lever pressing after cue onset, and rats were required to complete at least 90% of the trials before advancing to the next stage of training. Rats completed at least one more FR-1 session with this time limit while untethered, and for all subsequent sessions they were tethered to the voltammetry head-stage. The remaining stages of training included another session of FR-1 with a 30 ± 10 s ITI, and then FR-4 with a 45 ± 15 s ITI. Choice trials were then introduced, and all subsequent sessions began with a "training block" of 24 forced trials (12 for each option in pseudo-randomized order) followed by seven blocks (56 trials) of four forced and four choice trials, as in the block structure used with the first two cohorts. Rats then progressed to the mixed-contingency decision-making task.

Experiment three: mixed-contingency decision-making task. As in prior training for this third group, rats had to make an uncued head entry into the centrally located food magazine after the 45 ± 15 s ITI period ended to trigger the simultaneous onset of the cue light(s) and extension of

the lever(s). The first 24 trials of each session were all forced trials, serving as a training block that was then followed by seven blocks (56 trials) of four forced and four choice trials, totaling in 80 trials per session. The side of high-value option was counterbalanced across rats but remained fixed across sessions while determining each rat's effort requirements. After one session requiring four presses for either option, the effort requirement for the high-value reward was increased to eight presses for the next session and then incremented by eight presses in each subsequent behavioral session until each individual rat switched to preferring the LL option. For voltammetry recording sessions, we then used the second highest effort level at which individual rats had previously preferred the HM option when recording the moderate-cost sessions (4-40 presses across rats) and the second lowest effort level at which rats instead preferred the LL option when recording the high-cost sessions (32-72 presses across rats), with the order counterbalanced across rats. For this third experiment, the number of HH-preferred high-cost sessions was insufficient for analysis.

Fast-scan cyclic voltammetry recording sessions. The chronically implanted carbon-fiber microelectrodes were connected to a head-mounted voltammetric amplifier for dopamine detection by fast-scan cyclic voltammetry as previously described (Clark et al., 2010). A potential of -0.4 V (versus the Ag/AgCl reference) was applied to the carbon fiber and ramped to +1.3 V and back at a rate of 400 V/s. This voltammetric scan was applied at a frequency of 60 Hz for ~40 min prior to recording the behavioral sessions and then at 10 Hz for ~20 min prior to and throughout the recording session. To confirm that electrodes were capable of detecting chemically verified dopamine, a series of unexpected food pellets were delivered before and after each recording session. The voltammetry data from a recording session were included in the

analysis only if the pre- and post-session pellet delivery elicited dopamine release whose cyclic voltammogram (electrochemical signature) achieved a high correlation (correlation coefficient $r^2 \geq 0.75$ by linear regression) with that of a dopamine standard. For data included in Experiment One, a pair of recorded sessions with counterbalanced lever side assignments was obtained for at least one behavioral condition (moderate-cost and/or LL-preferred high-cost sessions). Because we have previously reported the same pattern of greater dopamine transmission for the high-value option regardless of the lever assignment configuration (i.e., no evidence for direction-selective encoding by mesolimbic dopamine) in experiment one (Figure 2.3), for experiments two and three we included all recorded sessions meeting the electrochemical and behavioral criteria regardless of whether the counterbalanced pair was obtained.

Statistical analyses. Post-criterion choice proportions were normalized with the arcsine transformation and compared to indifference using two-tailed, one-sample t-tests in SPSS. Voltammetry data analysis was carried out using software written in LabView and Matlab. Following 2000-Hz low-pass filtering, dopamine was isolated from the background-subtracted (one second prior to cue onset) voltammetric signal using chemometric analysis (Heien et al., 2005) using a standard training set based on stimulated dopamine release detected by chronically implanted electrodes (Clark et al., 2010). Dopamine concentration was estimated based on the average post-implantation electrode sensitivity. Noise spikes >1.5 nA versus the immediately preceding and following time points were removed (Gan et al., 2010), and the data were smoothed using a 0.5-s moving average.

The discriminability of cue-evoked dopamine responses to the different trial types was analyzed at each time point using the area under the receiver operating characteristic curve (auROC), an approach from signal detection theory (Green and Swets, 1966). The high-value option (HM or HH) was always coded as the positive cases in comparisons with the LL option, and auROC values were not rectified around 0.5 (i.e., if the LL option had evoked a greater dopamine response, the auROC values would have been less than 0.5). Significant discriminability at each time point was determined using a random permutation test, shuffling the trial types and re-computing the auROC, and repeating this for 2000 permutations to generate a null distribution. After correcting for multiple comparisons across time using a suprathreshold cluster-correction technique (Nichols and Holmes, 2002; Buschman et al., 2012), all time points outside the 95% confidence interval were considered statistically significant. To further assess whether dopamine responses to each type of choice trial were more discriminable from the dissimilar forced trials than from the corresponding forced trials within each session type, these auROC values following cue onset were compared using Wilcoxon signed rank tests. For graphical display purposes, all auROC values were transformed to a dopamine discriminability index ranging from -1 to 1: discriminability index = $2 \times (\text{auROC} - 0.5)$.

Histological verification of recording site. Animals were anesthetized with ketamine (100 mg/kg) and xylazine (20 mg/kg), and the recording site was marked by passing a current (~70 μA) through the carbon-fiber microelectrode for 20 s to make a small electrolytic lesion. Animals were perfused transcardially with physiological saline and then with four-percent paraformaldehyde in phosphate-buffered saline, in which brains also were post-fixed following removal from the skull. Brains were sunk in 15% sucrose solution in PBS for 24 hours, 30%

sucrose for at least 72 hours, flash frozen in dry ice, sectioned coronally (30-60 μm) on a cryostat, mounted on slides, and stained with a 0.5% cresyl violet solution.

Results

Rats performed a mixed-contingency decision-making task as previously described (Hollon et al., 2014) while mesolimbic dopamine transmission was monitored with fast-scan cyclic voltammetry at chronically implanted carbon-fiber microelectrodes (Clark et al., 2010) in the nucleus accumbens core (Figures 2.1 and 3.3). Briefly, the behavioral task consisted of blocks of eight discrete trials each separated by a 45 ± 15 s intertrial interval (ITI), with four single-option forced trials (two for each option in pseudorandomized order) followed by four choice trials in which both options were available. In experiment one, each trial began with the simultaneous onset of the cue light(s) and corresponding lever(s); in experiment two, cue light onset preceded lever presentation by 5 s. In all conditions, one option yielded one food pellet for four lever presses, designated the low-value / low-effort (LL) option. In the moderate-cost condition, the alternative option yielded four pellets four eight presses (high value / medium effort, HM). In the high-cost condition, the alternative option also yielded four pellets but for a higher effort requirement (high value / high effort, HH) which was determined individually for each rat during initial behavioral training (see Methods) but remained constant for a each rat within a given session, ranging from 32-128 presses per trial across rats. A behavioral criterion was defined as 75% choice for the HM option in moderate-cost sessions and for the LL option in high-cost sessions within a sliding window of twelve consecutive choice trials. If the rat failed to reach

criterion for the LL option within a particular high-cost recording session but instead reached the opposite criterion for the HH option, then this session was classified as an HH-preferred high-cost session, resulting in three session types: moderate-cost, HH-preferred high-cost, and LL-preferred high-cost sessions. As previously reported (Hollon et al., 2014), after reaching the respective behavioral criterion in each session type, rats in experiments one and two continued to exhibit significant preferences for the HM option in moderate-cost sessions, for the HH option in HH-preferred high-cost sessions, and for the LL option in LL-preferred high-cost sessions. Moreover, on single-option forced trials, we found discriminably greater cue-evoked mesolimbic dopamine release to the option yielding the high-value outcome (HM or HH) regardless of whether animals preferred this option or the instead preferred the LL alternative (Hollon et al., 2014). Because the dopamine response to the unexpected presentation of a reward-predictive cue provides a readout of the associated cached value, this dissociation between behavioral preference and dopamine-associated cached values permitted us to examine which cached value is represented by cue-evoked dopamine release when animals face choices between concurrently available options under circumstances in which the preferred option either is or is not associated with the greatest cached value. Therefore, in contrast to previous studies whose conflicting conclusions differed only because of the relative magnitude of the dopamine response during choices for a non-preferred option (Morris et al., 2006; Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012), the current investigations afforded us the opportunity to asses choices for options with high or low cached values in instances when each was the preferred option.

In moderate-cost sessions, the cue-evoked mesolimbic dopamine response in choice trials in which the rats selected the preferred HM option did not differ from the response in HM forced

trials but was significantly greater than in LL forced trials (Figure 3.1*A*), and similar results were observed for HH choice trials in the HH-preferred high-cost sessions (Figure 3.1*B*). However, because the preferred option chosen *is* the option with the greatest cached value available in these conditions, this result on its own does not differentiate between the alternative hypotheses of whether cue-evoked dopamine release during concurrent choice trials signals the cached value of the chosen option (Morris et al., 2006) or the greatest cached value available (Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012). In contrast, the LL-preferred high-cost sessions provided the critical test of these competing hypotheses, as the cached value of the option chosen when rats selected this preferred LL option was smaller than the cached value of unchosen alternative on these trials, the non-preferred HH option. If cue-evoked dopamine release signals the cached value of the chosen option according to a SARSA algorithm, then the dopamine response on these choice trials would be identical to the response on the corresponding LL forced trials. If dopamine transmission instead reflects the greatest cached value available regardless of choice, as in Q-learning algorithms, then the response on these choice trials would be identical to the response on the HH forced trials even though the rats chose the preferred LL option on these trials. The results from these LL-preferred high-cost sessions clearly supported the former hypothesis. The cue-evoked dopamine response during the choices for the LL option were indistinguishable from the response on the corresponding LL forced trials and were significantly smaller than the response observed during forced trials for the non-preferred HH option (Figure 3.1*C*). These findings demonstrate that on choices for the preferred option, cue-evoked mesolimbic dopamine release encodes cached value of this chosen option rather than the greatest cached value available. Likewise, this coding scheme also was apparent in experiment two, in which a 5-s delay was imposed between cue onset and lever presentation (Figure 3.1*D*-

F), indicating that cue-evoked dopamine signals the cached value of the chosen option even before the animals were able to enact their decision by beginning to press the corresponding lever.

The above analysis differentiates between SARSA and Q-learning classes of TDRL algorithms, which update state-action values, with our results strongly favoring SARSA. However, an alternative possibility is that the cached value read out by cue-evoked dopamine release represents a state value rather than a state-action value. The state value at the onset of a choice trial is a running average of the previously selected options and therefore is weighted towards the preferred option (which is experienced more often). Therefore, on choices for preferred options when there is a strong preference, algorithms that update the state value (e.g., $V(s)$ in actor-critic models) would read out cached values that are quite similar to those from SARSA; however, we can differentiate between these alternatives by analyzing the choice trials where animals selected the non-preferred option. Specifically, for the SARSA algorithm, we would expect dopamine release on non-preferred choice trials to differ from that on preferred choice and forced trials but be indistinguishable from that on non-preferred forced trials. Alternatively, if cue-evoked dopamine encodes the state value, then the cached value read out by dopamine release should be the same regardless of which option is ultimately chosen. In all session types for both experiments, we found significant discriminability between cue-evoked mesolimbic dopamine release during choices for the preferred versus non-preferred option (Figure 3.2). Unlike for preferred choices, on non-preferred choices dopamine release was significantly different from forced trials for the preferred option and, instead, was indistinguishable from that for forced trials for the non-preferred option. That is, dopamine release on choice trials matched the forced-trial

counterpart to the option that was subsequently chosen regardless of the overall preference within the session, consistent with a SARSA TDRL algorithm.

A potential caveat of the similarity between mesolimbic dopamine release on choice trials and their corresponding forced trials observed in the first two experiments is that animals might be perceiving these trial types as the same because the rats were already present at one of the cue locations and the onset of that cue was treated akin to a forced trial. Indeed, animals' physical proximity to a given option at trial onset has been suggested to bias their choices toward this nearer option (Morrison and Nicola, 2014). To exclude the possibility that proximity underlies the similarity of dopamine release between corresponding forced and choice trials, we ran an additional cohort of animals in a third variant of this task in which the rats had to make a head entry into the centrally-located food magazine to initiate each trial following the ITI period. This design served to center the rat in the operant chamber so that it was equidistant from each lever and was always facing the food magazine at trial onset rather than already being present at a particular lever. Despite this requirement and differences in the behavioral training procedures for this experiment (see Methods), the results fully supported the conclusions of experiments one and two (Figure 3.4), ruling out the influence of biases due to animal positioning. Thus, across all three experiments, we consistently found that cue-evoked mesolimbic dopamine transmission on choice trials encoded the cached value of the chosen option.

Discussion

By signaling temporal-difference prediction errors in a manner consistent with TDRL algorithms, phasic dopamine is widely believed to update the cached values assigned to state-action pairs, and the unexpected presentation of a reward-predictive cue evokes a dopamine response that provides a readout of the associated cached value (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997). Although numerous experiments have capitalized on this feature of dopaminergic prediction-error signals to obtain a neural readout of the cached value when a single option is presented in isolation (Fiorillo et al., 2003; Morris et al., 2004; Tobler et al., 2005; Morris et al., 2006; Roesch et al., 2007; Kobayashi and Schultz, 2008; Bromberg-Martin and Hikosaka, 2009; Day et al., 2010; Gan et al., 2010; Nasrallah et al., 2011; Sugam et al., 2012; Pasquereau and Turner, 2013; Hollon et al., 2014; Lak et al., 2014), it has remained controversial which cached value is represented when animals face choices between concurrently available options (Morris et al., 2006; Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012), resulting in conflicting conclusions regarding which variant of TDRL this neural system instantiates (Morris et al., 2006; Niv et al., 2006; Daw, 2007; Roesch et al., 2007). Here, we have demonstrated that cue-evoked mesolimbic dopamine encodes the cached value of the chosen option in a manner consistent with the "on-policy" state-action values of SARSA TDRL algorithms rather than "off-policy" Q-learning. Based on our recent demonstration that animals' preferences do not always align with the rank ordering of dopamine-reported cached values (Hollon et al., 2014), we were able to differentiate between SARSA and Q-learning by analyzing choices for the preferred option under circumstances in which this preferred option was not associated with the greatest cached value. The dopamine responses during these choice trials for

the preferred option were not discriminable at any time point from those in forced trials for the preferred option, but significantly diverged from those in forced trials for the non-preferred option. This approach provided a distinct advantage over the few previous studies addressing this question (Morris et al., 2006; Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012), as any conclusions drawn from these studies to differentiate between the TDRL algorithms depended entirely on their analysis of relatively infrequent choices for the option that on average was not preferred by the animals. Our analyses also underscore the importance of examining the full time course of dopamine dynamics beyond just the single timepoint at the peak of the response. Furthermore, despite the existing discrepancy in the literature (Morris et al., 2006; Niv et al., 2006; Daw, 2007; Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012), we observed remarkable internal consistency within the current work supporting this coding scheme, which replicated across all conditions in three distinct variants of our mixed-contingency decision-making task. Indeed, even when we subsequently analyzed dopamine transmission during the sparse non-preferred choices, cue-evoked dopamine transmission was significantly discriminable between choices for the preferred versus non-preferred options in all conditions. Collectively, our findings provide robust evidence that reward prediction errors encoded by mesolimbic dopamine are fully consistent with the SARSA coding scheme.

Although there were numerous differences between the behavioral task designs of the previous studies suggesting that cue-evoked dopamine responses either report the cached value of the subsequently chosen action (Morris et al., 2006) as predicted by SARSA (Rummery and Niranjan, 1994; Niv et al., 2006) or instead signal the greatest cached value available regardless of future choice (Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012) consistent with Q-

learning (Watkins, 1989; Daw, 2007), these divergent results often have been attributed to differences in species or recording location (Daw, 2007; Roesch et al., 2007; Morris et al., 2010; Morita et al., 2013; Morita, 2014; Morita and Kato, 2014). Specifically, the findings by Morris et al. (2006) supporting the SARSA algorithm came from dopamine neurons in the SNc of macaque monkeys, whereas those consistent with Q-learning were obtained in rats by recording the activity of dopamine neurons in the VTA (Roesch et al., 2007) or dopamine release in these mesolimbic neurons' terminal field within the nucleus accumbens core (Day et al., 2010; Sugam et al., 2012). It is perhaps surprising, therefore, that our recordings of mesolimbic dopamine in the current study align more closely with the findings from nigral dopamine neurons in macaques (Morris et al., 2006) rather than with mesolimbic dopamine transmission in rats (Roesch et al., 2007; Day et al., 2010; Sugam et al., 2012). However, because the conclusions of these previous studies depended on the relative magnitude of dopamine transmission during choices for the non-preferred option, it is unclear whether this alleged signaling of the greatest cached value during non-preferred choices truly implies the instantiation of a Q-learning algorithm. It is possible that this homogeneity across choice trials might be due to the animal making a mistake, such that this equivalently high dopamine response reflected the misrepresentation of the unchosen but preferred option's greater cached value. Alternatively, it also is possible that these choices for the non-preferred option instead were deliberate, exploratory decisions, and the resultant dopamine response may have incorporated an "exploration bonus" (Dayan and Sejnowski, 1996; Kakade and Dayan, 2002) that conceivably could have augmented this signal to a level comparable to that observed in preferred choice trials. Although the current experiments cannot resolve these various explanations for the previously observed magnitude of dopamine transmission during these anomalous choices for the non-preferred option, our ability to experimentally dissociate

animals' preferences from the rank ordering of dopamine-associated cached values (Hollon et al., 2014) obviated the need to analyze these non-preferred choice trials to reach our primary conclusion.

We note, however, that Roesch et al. (2007) did observe that immediately following the 500-ms odor-cue sampling period but still before the rats enacted a choice in their task, the VTA dopamine neuron firing rates rapidly diverged to signal the value of the subsequently chosen option, which is congruent with the finding of Morris et al. (2006) and the current study. Indeed, a subsequent study from this group using the same behavioral task (Takahashi et al., 2011) emphasized this encoding of the chosen option by dopamine neurons during choice trials as a key result, no longer highlighting the lack of discriminability within the first 500 ms after cue onset. Moreover, from the computational models employed in this work, these authors contended that SARSA was fully able to simulate their empirical results whereas Q-learning was not. Nevertheless, the original work by Roesch et al. (2007) is still frequently cited for its headline conclusion that dopamine neurons signal the greatest cached value available regardless of subsequent choice, a premise that has recently been used as justification for modeling dopamine neuron activity with a Q-learning algorithm (Morita et al., 2013; Morita, 2014; Morita and Kato, 2014). The results of the current study, in conjunction with this reinterpretation of the data from Roesch et al. (2007; Takahashi et al., 2011), instead lend stronger support for mesolimbic dopamine enacting a SARSA algorithm, as originally observed in the activity of nigral dopamine neurons (Morris et al., 2006).

In contrast to this evidence supporting the notion that dopaminergic prediction-error signals enact a SARSA-like algorithm, the two remaining studies (Day et al., 2010; Sugam et al., 2012), both from the Carelli lab, did not report any significant divergence in the cue-evoked dopamine responses according to the animals' subsequent choice. In light of the concern raised above regarding why animals selected their non-preferred option on a subset of choice trials and why the dopamine response during these anomalous choices was comparable to that during choices for the preferred option, a notable feature of the behavioral training undergone by animals in these studies was their degree of extended training under highly stable response contingencies. Specifically, the levers to which the task contingencies were assigned remained fixed across at least fourteen consecutive sessions of training, such that the side of the preferred option was constant throughout all training in these studies. In contrast, all the studies in which dopamine signaling adhered to a SARSA-like algorithm involved a greater degree of within-task dynamics. In the task used by Morris et al. (2006), the cues associated with the different reward probabilities were presented on either side of the screen in different trials and were selected by a left or right button press, such that the preferred option might be associated with either location and action in any given choice trial; the task used by Roesch et al. (2007; Takahashi et al., 2011) included multiple contingency changes such that the side of the preferred option reversed three times within each session; and in the current study, we recorded counterbalanced pairs of sessions in which the lever side assignments were reversed between sessions, separated by one to three sessions depending on the experimental cohort (see Methods). Intuitively, if anything, animals performing these more dynamic tasks might seem either more susceptible to making mistakes or more likely to make exploratory decisions, and yet these were the tasks in which cue-evoked dopamine was found to reliably encode the cached value of the chosen option

(Morris et al., 2006; Roesch et al., 2007; Takahashi et al., 2011; and the current results), whereas the two studies with protracted training under stable conditions yielded apparent homogeneity in cue-evoked dopamine across preferred and anomalous choices (Day et al., 2010; Sugam et al., 2012). Notably, a more recent study from the Carelli lab employing a task with dynamic, within-session contingency changes now also has found that mesolimbic dopamine release during concurrent choice trials reports the cached value of subsequently chosen option rather than the greatest cached value available (Saddoris et al., 2014). Therefore, the majority of these studies, including the three internally consistent replications within the current work, support the notion that cue-evoked dopamine signals the cached value of the chosen option according to a SARSA-like TDRL algorithm, whereas the two counterexamples in which dopamine transmission was more homogenous across choices for either option involved tasks with more protracted extended training in unchanging environments. Although a transition within this dopamine system from the SARSA-like signaling of chosen cached values to a Q-learning-like signaling of the greatest cached value regardless of choice would be surprising, this conjecture provides a tractable hypothesis for future work and may represent a more parsimonious explanation than do speculations about species or recording site differences to account for the results observed across this collection of studies.

Under most circumstances examined, therefore, the temporal-difference prediction errors encoded by dopamine transmission reveal the instantiation of an on-policy SARSA-like TDRL algorithm. By signaling the cached value of the chosen option, cue-evoked dopamine represents post-decision information, reflecting the value of a decision that had already been made by other neural structures (Morris et al., 2006; Niv et al., 2006). Interestingly, lesions of the ipsilateral

orbitofrontal cortex disrupt this signaling of the chosen cached value by dopamine neurons (Takahashi et al., 2011), supporting the notion that such cortical regions provide dopamine neurons with information about which option is chosen and the resulting dopamine transmission conveys changes in expectation of future reward based on the anticipated outcome of these choices.

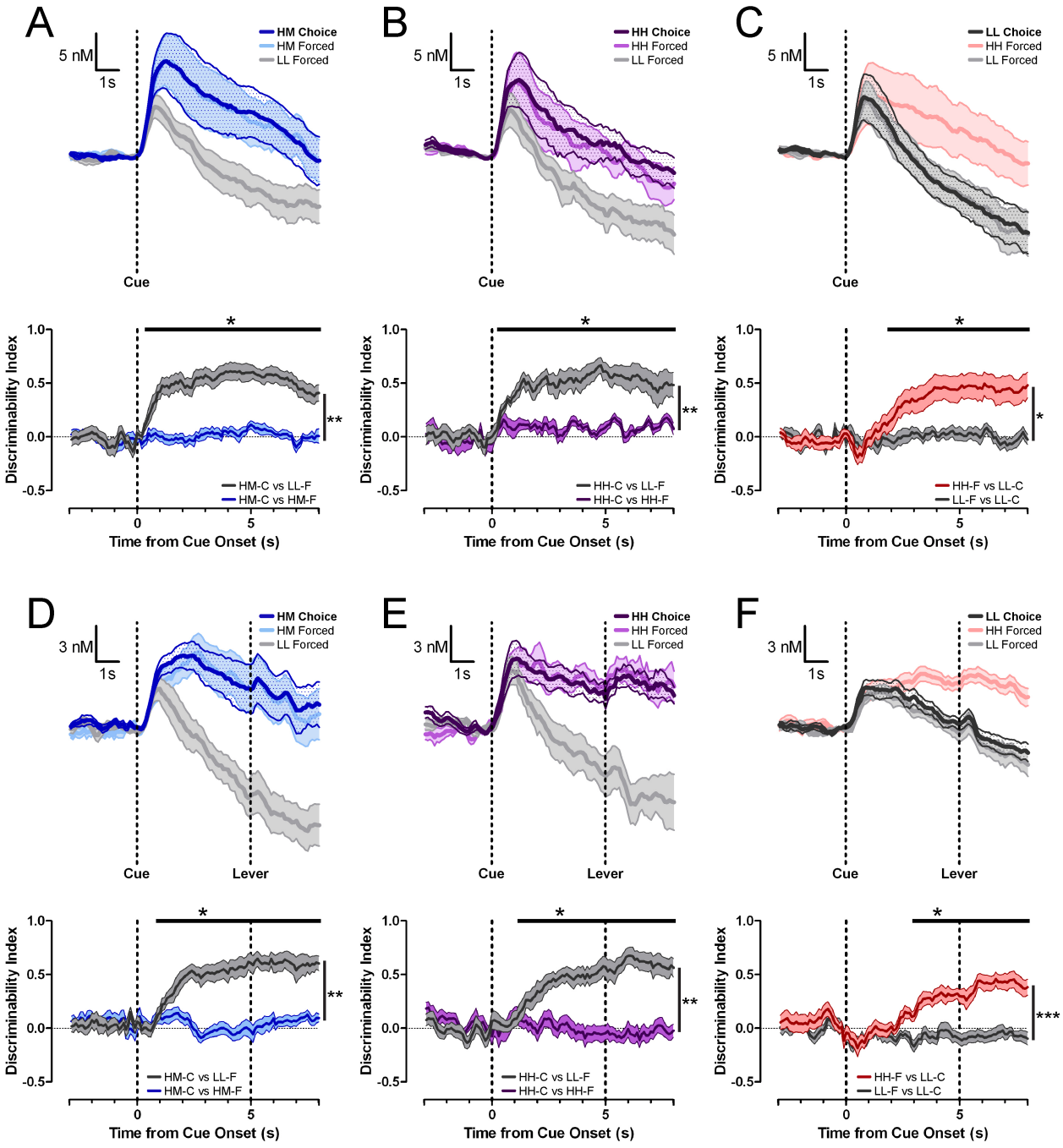

Figure 3.1. Voltammetry results from choices for the preferred option in experiments one and two.

Mean (± SEM) cue-evoked dopamine release during choices for the preferred in each session type from experiment one (*A*-*C*) and experiment two (*D*-*F*). The HM option was preferred in moderate-cost sessions (*A*, $n = 11$ recording sites; *D*, $n = 9$ recording sites), the HH option was

preferred in HH-preferred high-cost sessions (*B*, $n = 11$ recording sites; *E*, $n = 11$ recording sites), and the LL option was preferred in LL-preferred high-cost session (*C*, $n = 10$ recording sites; *F*, $n = 15$ recording sites). Choice trial traces are overlayed on forced trial responses (Hollon et al., 2014). Below: mean (± SEM) discriminability index time series comparing choices for the preferred option to each forced trial type. Horizontal bars indicate time points of significant discriminability for preferred choice trials vs. non-preferred forced trials (*$P < 0.05$, permutation tests); the comparisons of preferred choice and forced trials never reached significance. Vertical bars indicate significant differences between the two discriminability index traces following cue onset (* $p < 0.05$, ** $p < 0.005$, *** $p < 0.001$, Wilcoxon signed rank test).

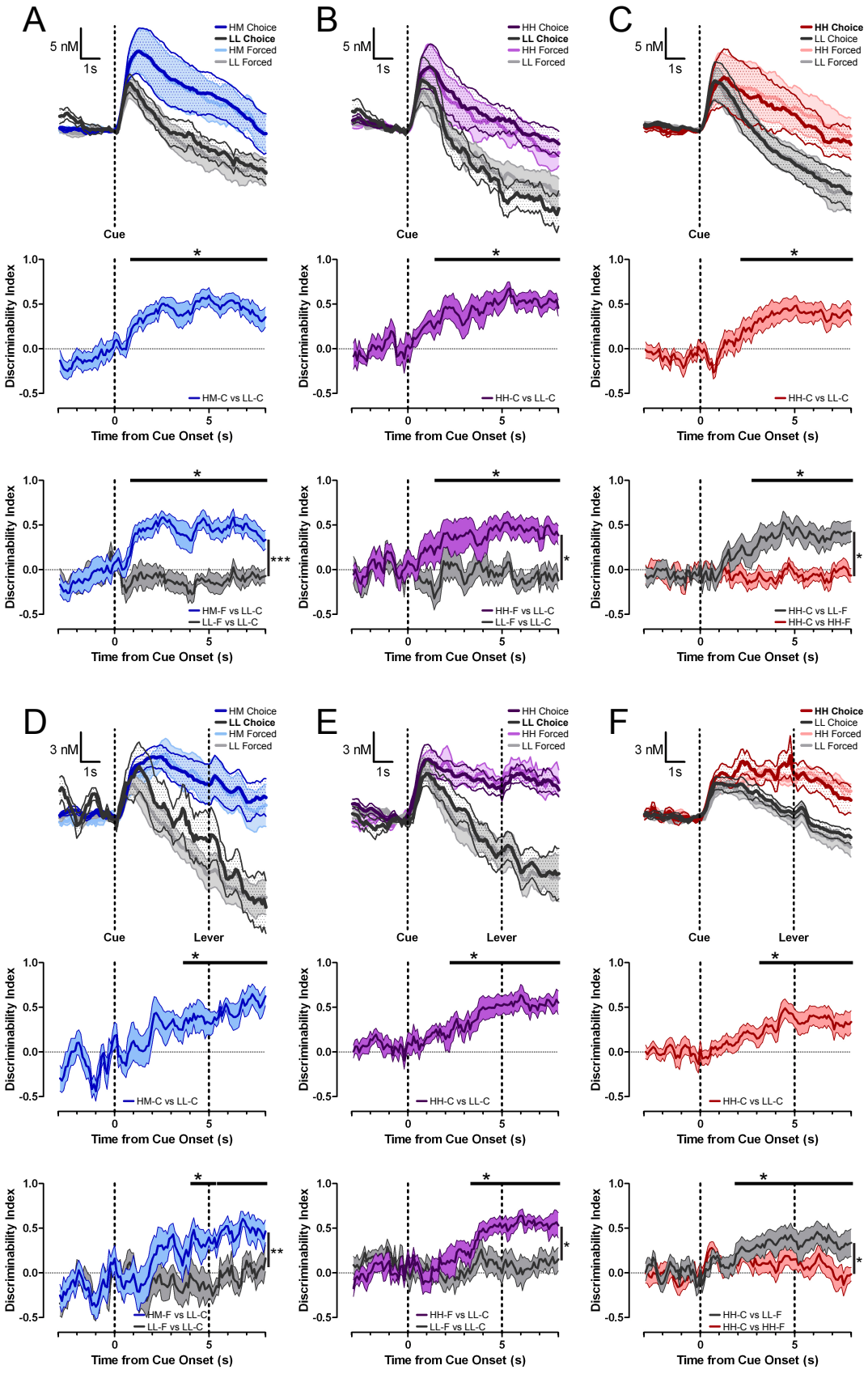
A
B
C
D
E
F
5 nM
3 nM
1s
HM Choice
LL Choice
HM Forced
LL Forced
HH Choice
HH Forced
Cue
Lever
Discriminability Index
Time from Cue Onset (s)
HM-C vs LL-C
HH-C vs LL-C
HM-F vs LL-C
LL-F vs LL-C
HH-F vs LL-C
HH-C vs LL-F
HH-C vs HH-F

Figure 3.2. Voltammetry results from choices for the non-preferred option in experiments one and two.

Mean (± SEM) cue-evoked dopamine release during choices for the non-preferred in each session type from experiment one (*A*-*C*) and experiment two (*D*-*F*). The LL option was not preferred in moderate-cost sessions (*A*, $n = 11$ recording sites; *D*, $n = 8$ recording sites) and in HH-preferred high-cost sessions (*B*, $n = 8$ recording sites; *E*, $n = 9$ recording sites), and the HH option was not preferred in LL-preferred high-cost session (*C*, $n = 10$ recording sites; *F*, $n = 13$ recording sites). Choice trial traces are overlayed on forced trial responses (Hollon et al., 2014) and preferred choice trial responses from Figure 3.1. The *n* for certain session types is lower than in Figure 3.1 because not all recorded sessions contributed choices for the non-preferred option (i.e., if the rat chose the preferred option in 100% of post-criterion choice trials). Below, middle row: mean (± SEM) discriminability index time series comparing choices for the preferred vs. non-preferred option. Horizontal bars indicate time points of significant discriminability (*$p < 0.05$, permutation tests). Below, lower row: mean (± SEM) discriminability index time series comparing choices for the non-preferred option to each forced trial type. Horizontal bars indicate time points of significant discriminability for non-preferred choice trials vs. preferred forced trials (* $p < 0.05$, permutation tests); the comparisons of non-preferred choice and forced trials never reached significance. Vertical bars indicate significant differences between the two discriminability index traces following cue onset (* $p < 0.05$, ** $p < 0.005$, *** $p < 0.001$, Wilcoxon signed rank test).

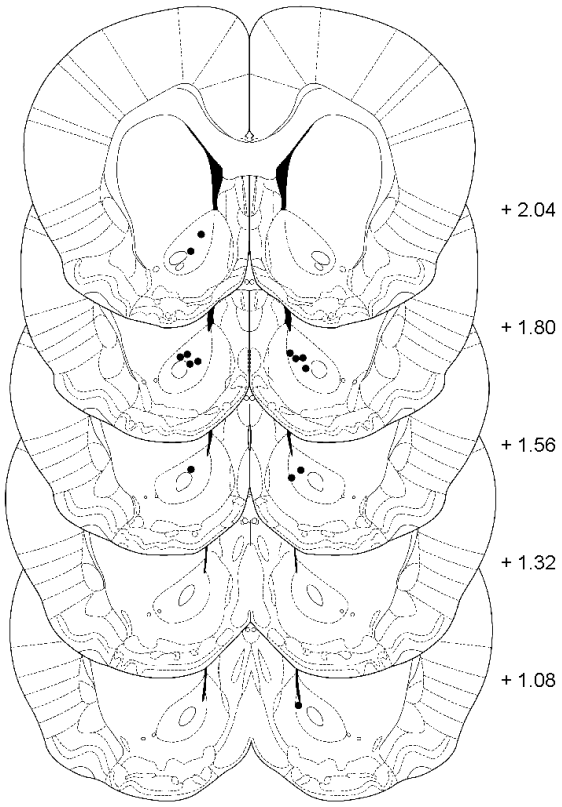

Figure 3.3. Recording locations in the nucleus accumbens core for experiment three.

Electrode placements for rats in experiment three (placements for rats in experiments one and two are shown in Fig. 2.1). The numbers next to each section indicate distance in mm anterior to bregma. Adapted from the atlas of Paxinos and Watson (2005).

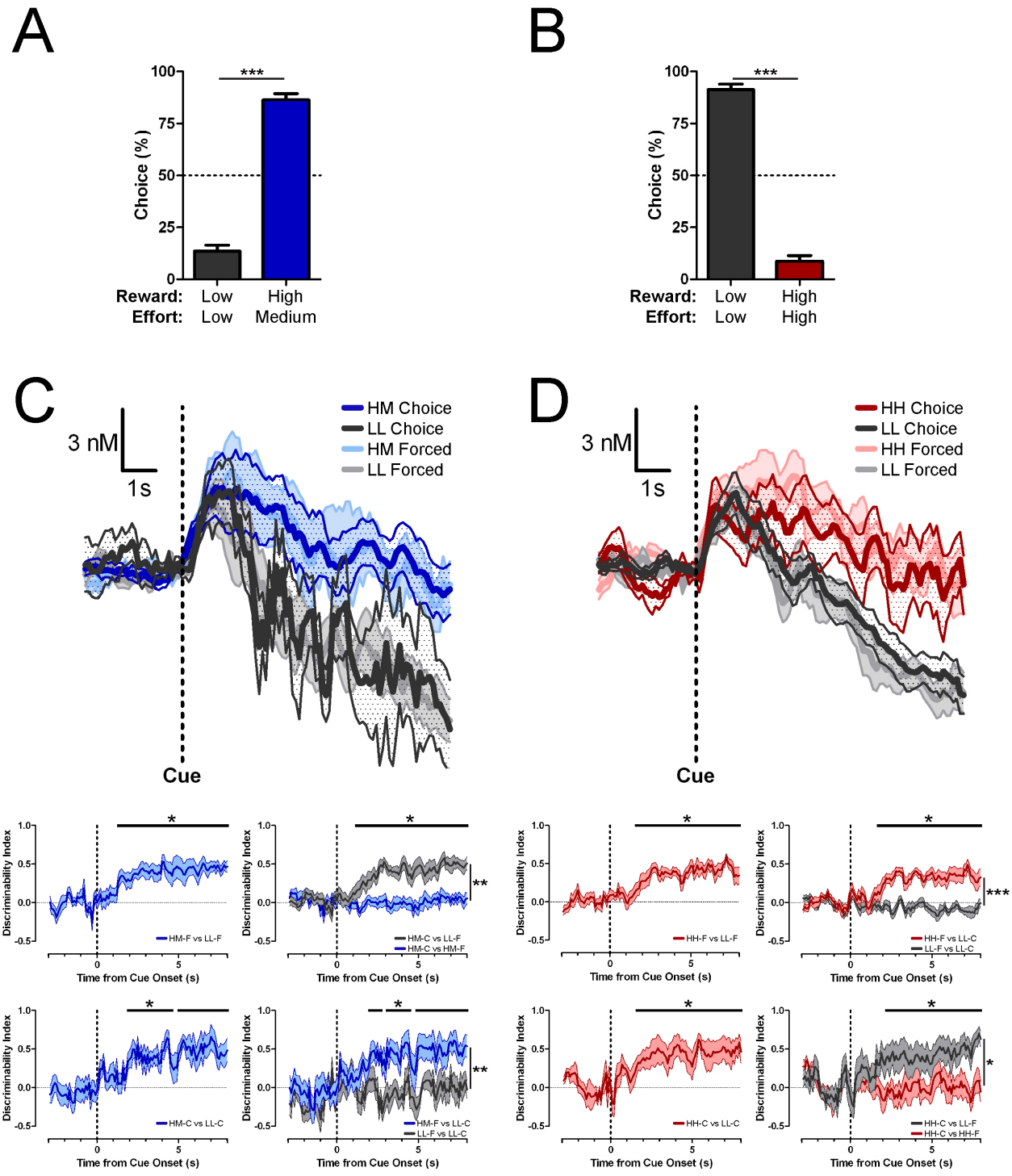
A
B
C
D
Choice (%)
Reward:
Effort:
Low
High
Medium
***
HM Choice
LL Choice
HM Forced
LL Forced
HH Choice
HH Forced
LL Forced
3 nM
1s
Cue
Discriminability Index
Time from Cue Onset (s)
HM-F vs LL-F
HM-C vs LL-F
HM-C vs HM-F
HH-F vs LL-F
HH-F vs LL-C
LL-F vs LL-C
HM-C vs LL-C
HM-F vs LL-C
HH-C vs LL-C
HH-C vs LL-F
HH-C vs HH-F

Figure 3.4. Behavioral and voltammetry results from experiment three, in which rats were required to make a centering head-entry to initiate trials.

(*A*) Mean (+ SEM) post-criterion percent choice from moderate-cost sessions ($t_9 = 9.979$, *** $P = 3.64 \times 10^{-6}$, $n = 10$ rats) and LL-preferred high-cost sessions ($t_9 = 11.99$, *** $P = 7.78 \times 10^{-7}$, $n = 10$ rats). (*B*) Mean (± SEM) cue-evoked dopamine release during all trial types in moderate-cost sessions ($n = 12$ recording sites, 11 from sessions contributing non-preferred LL choices). HM (blue) and LL (gray) choice trials are overlayed on HM (light blue) and LL (light gray) forced trials. Below: mean (± SEM) discriminability index time series for each comparison. Top left: HM forced vs. LL forced. Bottom left: HM choice vs. LL choice. Top right: HM choice vs. LL forced (gray, with significant time points indicated by the horizontal bar, * $P < 0.05$, permutation test) and HM choice vs. HH choice (blue, never reached significance). Vertical bar indicating significant difference between the two discriminability index traces following cue onset ($Z = -2.981$, ** $P = 0.003$, Wilcoxon signed rank test). Bottom right: HM forced vs. LL choice (blue, with significant time points indicated by the horizontal bar, * $P < 0.05$, permutation test) and LL forced vs. LL choice (gray, never reached significance). Vertical bar indicating significant difference between the two discriminability index traces following cue onset ($Z = -2.934$, ** $P = 0.003$, Wilcoxon signed rank test). (*C*) Mean (± SEM) cue-evoked dopamine release during all trial types in high-cost sessions ($n = 13$ recording sites, 7 from sessions contributing non-preferred HH choices). HH (red) and LL (gray) choice trials are overlayed on HH (light red) and LL (light gray) forced trials. Below: mean (± SEM) discriminability index time series for each comparison. Top left: HH forced vs. LL forced. Bottom left: HH choice vs. LL choice. Top right: HH forced vs. LL choice (red, with significant time points indicated by the horizontal bar, * $P < 0.05$, permutation test) and LL forced vs. LL choice (gray, never reached significance). Vertical bar indicating significant difference between the two discriminability index traces following cue onset ($Z = -3.180$, *** $P = 0.001$, Wilcoxon signed rank test). Bottom right: HH choice vs. LL forced (gray, with significant time points indicated by the horizontal bar, * $P < 0.05$, permutation test) and HM choice vs. HH choice (red, never reached significance). Vertical bar indicating significant difference between the two discriminability index traces following cue onset ($Z = -2.366$, * $P = 0.018$, Wilcoxon signed rank test).

Chapter 4

Discussion

The results presented in this dissertation have demonstrated that the temporal-difference prediction errors signaled by mesolimbic dopamine transmission enact an on-policy SARSA-like algorithm reporting the cached value of the chosen option, but the cached values learned within this system are not always consistent with animals' behavioral preferences and therefore are not sufficient as the sole basis upon which these decisions are made. Therefore, although the findings of Chapter 3 address and clarify a longstanding algorithmic discrepancy in the literature, the findings of Chapter 2 have more profound implications that challenge a fundamental premise of current neuroeconomic theories of decision making. Specifically, the prevailing hypothesis, that the cached state-action values learned and read out through dopaminergic prediction-error signals are equivalent to the final subjective values providing the sole basis for determining decisions, simply cannot hold in a domain-general manner. This work raises important unresolved questions regarding what else is needed if these dopamine-associated cached values do still contribute to decision making, and what roles both dopamine and its associated cached values might actually play in this process.

4.1) The parts missing from contemporary theories

The prevailing accounts of the neural basis of value-guided decision making attest that dopaminergic prediction errors provide a teaching signal for learning the subjective value of state-action pairs and that the comparison of these cached values alone is sufficient for determining action selection. The current work does not challenge the premise that phasic dopamine transmission meets the basic criteria for encoding a temporal-difference-prediction-

error signal. There is ample evidence in the literature suggesting that the activity of many recorded dopamine neurons and patterns of terminal release indeed resemble this signal in a quantitative manner (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997; Waelti et al., 2001; Fiorillo et al., 2003; Morris et al., 2004; Bayer and Glimcher, 2005; Hart et al., 2014), and dopaminergic prediction errors can causally influence conditioned approach and reinforce instrumental actions in a manner predicted by the computational accounts (Tsai et al., 2009; Witten et al., 2011; Kim et al., 2012; Rossi et al., 2013; Steinberg et al., 2013; Ilango et al., 2014a; 2014b; Steinberg et al., 2014). However, although these signals do incorporate several subjective attributes beyond objective expected value that also affect animals' preferences (Roesch et al., 2007; Kobayashi and Schultz, 2008; Bromberg-Martin and Hikosaka, 2009; Day et al., 2010; Nomoto et al., 2010; Nasrallah et al., 2011; Sugam et al., 2012; Lak et al., 2014; Stauffer et al., 2014), they do not do so in a universal and domain-general manner, as cost-related attributes such as effort and aversiveness robustly influence behavioral preference yet they are weakly or inconsistently encoded by dopamine transmission (Ravel and Richmond, 2006; Gan et al., 2010; Wanat et al., 2010; Fiorillo, 2013; Pasquereau and Turner, 2013; Hollon et al., 2014). Consequently, there are circumstances in which the rank ordering of dopamine-associated cached values are diametrically opposed to the ordinal utility of the available options, as demonstrated in Chapter 2 (Hollon et al., 2014). This result illustrates a violation of the prevailing account's position that dopamine-associated cached values are equivalent to the subjective values providing the sole basis for comparing and selecting between available options.

Given that these dopamine-associated cached values alone are not sufficient as the basis for economic decisions, what else is needed if they still are to contribute to these decisions? At

minimum, there must be a separate representation of costs, and if these cost representations are incorporated into a final domain-general subjective value used for decision making, then this incorporation must occur downstream of the dopamine-associated cached values. It remains unknown whether there are common neural representations of all cost-related attributes (e.g., effortful response costs, aversiveness, and other potentially unidentified costs) or whether separate neural substrates underlie each, but there is some evidence that such costs may be represented by subsets of neurons within regions such as the anterior cingulate cortex (ACC) (Walton et al., 2003; Rudebeck et al., 2006; Hillman and Bilkey, 2010; Kennerley et al., 2011; Amemori and Graybiel, 2012; Koga et al., 2015), anterior insular cortex (Prévost et al., 2010; Palminteri et al., 2012; Skvortsova et al., 2014), and basolateral amygdala (Ghods-Sharifi et al., 2009; McHugh et al., 2014). Notably, each of these regions sends dense projections to the striatum. However, the stipulation that such cost information should be incorporated downstream of the dopamine-associated cached values (and likewise, not feed back onto the majority of dopamine neurons in a manner that would cause such costs to be incorporated into their activity), provides constraints on the functional anatomy which, although admittedly speculative, can help generate specific predictions for future research. For example, there have been suggestions that whereas striatal MSNs of the canonical direct and indirect basal ganglia pathways contribute to action selection as described in the introduction (section *1.1c*), a distinct system of projections originating from striosomal MSNs in the striatal patch compartment may be responsible for evaluative feedback (Stephenson-Jones et al., 2013), and only these patch compartment MSNs synapse directly onto midbrain dopamine neurons (Watabe-Uchida et al., 2012). Therefore, one might predict that the subpopulation of cortical neurons specifically encoding costs would not synapse onto these patch compartment MSNs but could selectively influence the canonical

pathways biasing action selection. Gross anatomical tracing data ostensibly seems at odds with this prediction, however, as the ACC sends projections to the striatal patch compartment that are denser than those of any other frontal area examined other than the posterior OFC (Eblen and Graybiel, 1995). Nevertheless, these ACC tracings also revealed inputs to the surrounding extrastriosomal matrix, leaving open the possibility that these particular projections originate from the subpopulation of ACC neurons encoding effortful response costs. Such specific predictions could be tested using recently developed mouse lines permitting genetic targeting of patch versus matrix MSNs (Gerfen et al., 2013) in combination with monosynaptic, retrogradely transported viruses (Wickersham et al., 2007; Wall et al., 2010) and optogenetic cell-type identification during electrophysiological recording (Lima et al., 2009; Jin and Costa, 2010; Cohen et al., 2012; Jin et al., 2014). As an aside, it is also notable that the OFC, the other cortical region exhibiting particularly dense projections to the striatal patch compartment, has been found to contain fewer neurons encoding effort costs compared to the populations encoding other economic attributes (Kennerley et al., 2009; Kennerley et al., 2011), and lesions of this area do not disrupt effort-based decision making (Rudebeck et al., 2006). The OFC will be discussed further in the context of choices within "goods space" and multiple valuation systems in sections below.

4.2) Amending the algorithms

At the algorithmic level, the proposed downstream incorporation of cost-related attributes requires amendments to the dominant models of action selection, in which the cached values

learned through iterative updating by temporal-difference prediction errors provide the sole basis for action selection and do not currently include a separate representation of costs. As an example of a possible step in the right direction, a recent study modeled reward value maximization and effort cost minimization as distinct learning processes, with each using an update rule based on prediction errors for their respective dimensions (Skvortsova et al., 2014). In their functional magnetic resonance imaging results, they observed blood-oxygen-level-dependent activation correlating with reward-prediction errors in the striatum whereas effort-related correlates were observed in the ACC and anterior insula. Note that the proposed downstream integration of cost representations not incorporated into the dopamine-associated cached values is distinct from the earlier proposal of opponent systems responsible for signaling positive versus negative prediction errors (Daw et al., 2002). That opponency model proposed that the dopamine-signaled positive prediction errors were combined with a separate representation of negative prediction errors into a single term for updating the same cached value, indicative of upstream integration that would be reflected in subsequent cue-evoked prediction errors reporting this cached value; this prediction of the opponency model (Daw et al., 2002) is inconsistent with the results presented in Chapter 2 (Hollon et al., 2014).

A more recently proposed "opponent actor" model instead learns separate weights for corticostriatal synapses onto direct and indirect pathway neurons (Collins and Frank, 2014), and its authors claim that segregating effort costs into the weights of the latter pathway and differentially combining the weights of each pathway depending on tonic dopamine levels allows this model to account for many previously observed results, including both the differential sensitivity to learning about gains versus losses depending tonic dopamine concentration (Frank

et al., 2004) and the influence of dopamine receptor antagonists on effort-based decision making (Salamone et al., 2007). However, this model uses the same dopaminergic prediction-error term for learning each set of weights, so it is not apparent how this model would permit learning to avoid high-effort options using only a teaching signal that itself is relatively insensitive to effort. Nevertheless, this computational architecture may be amenable to modifications that might better accommodate the findings of the current experiments. This model's consideration of both learning and performance effects makes this approach an appealing combination of traditional TDRL algorithms and other models of instrumental vigor based on tonic dopamine concentration (Niv et al., 2007; Dayan, 2012). Although traditional voltammetric data acquisition as performed in the current experiments depends on background subtraction and therefore is not well-suited for the estimation of absolute dopamine tone, a recently developed insight permits the determination of tonic concentration within a manner of seconds using the same FSCV instrumentation: by simply replacing the standard 10-Hz triangle waveform application (-0.4 V to 1.3 V an back at 400 V/s) with the constant -0.4 V holding potential for several seconds and then reverting to the 10 Hz waveform for measurement, tonic dopamine concentration can be estimated from this controlled adsorption period (Atcherley et al., 2013; Atcherley et al., 2015). This innovation should make it more feasible in future research to test hypotheses relating behavior to tonic dopamine concentrations in addition to phasic changes typically detected with FSCV.

Elaborating on the standard TDRL framework for learning and action selection, a particularly intriguing model uses a hierarchical reinforcement learning algorithm to account for the selection of and persistence in behavior requiring the organism to overcome effortful response costs to

obtain reward (Holroyd and McClure, 2015). In hierarchical reinforcement learning, a lower level of the algorithm learns the values of "primitive" actions and selects actions based on these cached values as in standard ("flat") TDRL, and a higher level(s) learns the values of superordinate options or abstract goals defined as collections or sequences of primitive actions (Sutton et al., 1999; Botvinick et al., 2009). In this particular model (Holroyd and McClure, 2015), the ACC is proposed to function as the higher-level actor responsible for selecting between these overarching options and providing control over the striatal cells responsible for the selection and execution of the primitive action elements. Importantly, the cached state-action values learned within the striatum through dopaminergic prediction-errors explicitly do not incorporate effortful response costs; these costs are subtracted from the dopamine-associated cached values in a subsequent step before being compared for action selection. Therefore, while the ACC provides a hierarchical control level that can bias an organism toward selecting options requiring greater effort to obtain greater reward, the effort costs never directly enter into the primitive striatal state-action values or their associated dopaminergic prediction-error signals. To my knowledge, this computational algorithm may be the most congruent with the results of the current experiments (Hollon et al., 2014), and this hierarchical reinforcement learning framework may represent a more viable approach for modeling the neurobiology of value-guided decision making than do the standard algorithms proposed within the current prevailing neuroeconomic accounts.

4.3) Selecting actions versus choosing goods

This notion of hierarchical levels of abstraction also pertains to the current debate in the field regarding whether decisions are made in the space of goods versus actions. As detailed in the introduction to this dissertation, the dominant view in neuroeconomics maintains that while action-independent (abstract) subjective values may be represented by neurons in structures such as the OFC, these values ultimately serve to bias decisions made in "action space" through a winner-take-all competition between effector-specific populations of sensorimotor neurons (Platt and Glimcher, 1999; Redgrave et al., 1999; Dorris and Glimcher, 2004; Glimcher et al., 2005; Samejima et al., 2005; Lau and Glimcher, 2007, 2008; Louie and Glimcher, 2010; Seo et al., 2012). In contrast, proponents of a "goods-based" model argue that organisms first select the good or option they want and only subsequently prepare and perform an action to obtain that good (Padoa-Schioppa, 2011; Cai and Padoa-Schioppa, 2014). Although a full discussion of the evidence supporting each side of this debate is beyond the scope of this dissertation, the current results may bear on how dopamine-associated cached values might fit into either scheme. As discussed throughout, if decisions are made in the space of actions with action-specific subjective value representations, then cost-related attributes must be combined downstream with the dopamine-associated cached values if these cached values are to contribute to action selection. In this regard, these cached values could be more aptly described as conveying the value of the reward outcome independent of the cost from actions required to obtain it, which ostensibly sounds more like the value of a "good" or goal. However, the original formulation of the good-based account explicitly states that the subjective values used for deciding between goods also incorporate costs into a domain-general common currency represented in the OFC

(Padoa-Schioppa, 2011), although it is unclear how exactly the cost of actions might be computed and incorporated without some parallel representation of these actions as well (Cisek, 2012). Whereas subpopulations of OFC neurons do seem to represent information related to the subjective value of reward outcomes (Padoa-Schioppa and Assad, 2006), there currently is little evidence for the incorporation of effort costs into these representations, and OFC lesions have been found not to disrupt effort-based decision making (Rudebeck et al., 2006). Therefore, these OFC representations of the value of outcomes associated with stimuli, independent of response costs, may share this commonality with the dopamine-associated cached values. Although Padoa-Schioppa's good-based account is largely dismissive of these cached values (Padoa-Schioppa, 2011), there is ample anatomical and functional evidence for circuitry involving the OFC, ventral and dorsomedial striatum, and VTA in the encoding and use of related though non-redundant outcome-value information (Takahashi et al., 2009; Groenewegen and Uylings, 2010; Haber and Knutson, 2010; Sesack and Grace, 2010; Lodge, 2011; McDannald et al., 2011; Takahashi et al., 2011; Watabe-Uchida et al., 2012; Gremel and Costa, 2013; Stott and Redish, 2014). Another account of economic decision making, still compatible with the claim that certain decisions can be made based on abstract goals or goods (Wunderlich et al., 2010), also emphasizes OFC-based value representations but recognizes that these subjective stimulus-outcome values must be combined with independent representations of action costs prior to making choices (Rangel and Hare, 2010; Rangel and Clithero, 2014), consistent with our current claims regarding the possible role for dopamine-associated cached values in action selection. Therefore, whether economic decisions are ultimately made in the space of goods or of actions, either framework would require a more restricted definition of these dopamine-associated cached values than the domain-general subjective values with which they are often equated.

Nevertheless, there may be greater precedent for at least some researchers discussing a valuation system pertaining to the value of goals or stimulus outcomes separate from the action costs required to obtain them (Rangel and Hare, 2010; Rangel and Clithero, 2014).

4.4) Multiple valuation systems and the roles of dopamine

Another dichotomy that has gained increasing attention within neuroeconomics, distinct though perhaps not entirely orthogonal to the consideration of choices occurring in goods- versus action-space, pertains to the existence of multiple valuation systems that compete for control of animals' behavior. Many variants of such multiple-systems accounts have been discussed throughout the history of the multiple subfields upon which neuroeconomics has been built (Tolman, 1949; Dickinson, 1985; Colwill and Rescorla, 1986; Balleine and Dickinson, 1998; Sutton and Barto, 1998; Toates, 1998; Kahneman, 2003; Poldrack and Packard, 2003), and the most recent formulation to gain prominence in the field is that of "model-free" versus "model-based" systems, as imported from the machine learning literature (Sutton and Barto, 1998). Model-free systems are comprised of the TDRL algorithms described in the introduction (section *1.1b*) for learning cached values through direct experience and selecting actions based on these cached values, whereas model-based systems learn additional information about the environment such as the state transition probabilities and the actual identities of reward outcomes. As we (Clark et al., 2012) and many others have extensively reviewed elsewhere (Daw et al., 2005; Rangel et al., 2008; Redish et al., 2008; Frank, 2011; Ito and Doya, 2011; Lee et al., 2012; Dolan and Dayan, 2013; Daw and O'Doherty, 2014), the model-free valuation system is thought to

underlie more stimulus-driven and inflexible behavior such as habits, whereas model-based representations permit more deliberative and flexible planning and goal-directed behavior, though mechanisms underlying the arbitration between these putative systems remains poorly understood (Daw et al., 2005; Keramati et al., 2011; Lee et al., 2014). Here, I focus the discussion on how the results of the current experiments might bear on considerations of multiple valuation systems and the possible roles of dopamine therein.

Our demonstration that the dopamine-associated cached values, learned through a model-free SARSA-like TDRL algorithm, are not sufficient as the basis for action selection raises the possibility that these cached values are only used in certain circumstances whereas an alternative valuation system is used in other situations. Specifically, these cached values could still provide the basis for animals' choices when their preferences were predominantly driven by the differences in reward magnitude, an attribute robustly incorporated into the cached values, but an alternative system may have been used when effortful response costs overshadowed the tradeoff with reward size. The results of the current experiments do not speak to the precise nature of such an alternative valuation system (i.e., whether it necessarily conforms to the any of the various computational algorithms proposed for model-based learning systems), as these results primarily reveal the limitations of the supposed model-free cached values. Moreover, this suggestion, that organisms might use their cached value system in circumstances where we observed these cached values to positively correlate with their preferences but not in other situations where a negative correlation was observed, merely re-describes our results rather than providing a satisfactory account for when or why an organism might use one system as opposed to another.

It is entirely possible that if an alternative valuation system is available, whether model-based or otherwise, the animals plausibly could be using this alternative system for making all cost-benefit decisions, and contrary to the prevailing neuroeconomic theories, the dopamine-associated cached values may not be critical for action selection at all. Indeed, dopamine-deficient mice are still capable of learning and expressing preferences for sweet solutions over water (Cannon and Palmiter, 2003), and dopamine receptor antagonists do not alter the relative allocation of animals' behavior elicited by outcome-specific conditioned stimuli (Ostlund and Maidment, 2012). These perturbations of dopamine signaling, however, cause profound reductions in the frequency with which animals initiate reward-seeding behavior and the general invigoration of such behavior by reward-predictive stimuli. Such findings are consistent with the broader behavioral pharmacology literature providing evidence that dopamine itself is important for rapidly responding to reward-predictive cues (Robbins and Everitt, 2007; Nicola, 2010; du Hoffmann and Nicola, 2014) and overcoming effortful response costs (Salamone et al., 2007; Floresco et al., 2008; Salamone et al., 2009). My supervisors had previously suggested a central role for cue-evoked dopamine when animals assess whether potential benefits are worth the effort required to obtain them, with the magnitude of dopamine release theorized to set a cost-expenditure threshold depending on the anticipated reward value (Phillips et al., 2007). But perhaps even this proposal is bestowing too analytical a role onto dopamine; cue-evoked dopamine and the cached values it reads out may indeed be important for facilitating the execution or performance of an action once selected, but this role in behavioral activation or invigoration would be secondary to a separate system responsible for actually selecting which goal or action to pursue. Therefore, for a field with the primary objective of understanding how

organisms make value-guided decisions, neuroeconomics may have given too prominent a role to this one neuromodulator at the center of its most prevalent theories.

4.5) Concluding remarks

To the admittedly oversimplified assertion with which I opened my introduction – that "we now have all the moving parts" for understanding value-based decision making (Glimcher, 2009) – I would propose the following counter-assessment of our current state of knowledge: we may have identified some of these moving parts, and we may have begun to figure out what these parts can and cannot do, but we are still a long way off from fully understanding how they all work together, how our everyday choices emerge from these systems, and what exactly we should aim to alter when something goes awry. Hopefully by pursuing a deeper and more thorough understanding of the basic functions of those pieces we have identified so far, we have begun and will continue to gain a better sense of what is missing to help us determine where we need to go start looking next. As incomplete and unsatisfying as this more realistic assessment might be, at least we can take comfort in knowing that there are plenty of open questions left for us all to keep investigating.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transaction on Automatic Control 19:716-723.

Amemori K, Graybiel AM (2012) Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. Nat Neurosci 15:776-785.

Atcherley CW, Laude ND, Parent KL, Heien ML (2013) Fast-scan controlled-adsorption voltammetry for the quantification of absolute concentrations and adsorption dynamics. Langmuir 29:14885-14892.

Atcherley CW, Wood KM, Parent KL, Hashemi P, Heien ML (2015) The coaction of tonic and phasic dopamine dynamics. Chem Commun (Camb) 51:2235-2238.

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

Bartra O, McGuire JT, Kable JW (2013) The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. Neuroimage 76:412-427.

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129-141.

Bayer HM, Lau B, Glimcher PW (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. J Neurophysiol 98:1428-1439.

Bindra D (1974) A motivational view of learning, performance, and behavior modification. Psychol Rev 81:199-213.

Bolles RC (1972) Reinforcement, expectancy, and learning. Psychological Review 79:394-409.

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113:262-280.

Brischoux F, Chakraborty S, Brierley DI, Ungless MA (2009) Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proc Natl Acad Sci U S A 106:4894-4899.

Bromberg-Martin ES, Hikosaka O (2009) Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron 63:119-126.

Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010a) Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 68:815-834.

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010b) A pallidus-habenula-dopamine pathway signals inferred stimulus values. J Neurophysiol 104:1068-1076.

Brown HD, McCutcheon JE, Cone JJ, Ragozzino ME, Roitman MF (2011) Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. Eur J Neurosci 34:1997-2006.

Burnham KP, Anderson DR (2002) Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach. New York: Springer-Verlag.

Buschman TJ, Denovellis EL, Diogo C, Bullock D, Miller EK (2012) Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. Neuron 76:838-846.

Bush RR, Mosteller F (1951) A mathematical model for simple learning. Psychol Rev 58:313-323.

Cai X, Padoa-Schioppa C (2014) Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. Neuron 81:1140-1151.

Cai X, Kim S, Lee D (2011) Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. Neuron 69:170-182.

Cannon CM, Palmiter RD (2003) Reward without dopamine. J Neurosci 23:10827-10831.

Caplin A, Glimcher PW (2014) Basic methods for neoclassical economics. In: Neuroeconomics: Decision Making and the Brain, 2nd Edition (Glimcher PW, Fehr E, eds), pp 3-17. London: Elsevier Academic Press.

Chaudhury D et al. (2013) Rapid regulation of depression-related behaviours by control of midbrain dopamine neurons. Nature 493:532-536.

Cisek P (2012) Making decisions through a distributed consensus. Curr Opin Neurobiol 22:927-936.

Clark JJ, Sandberg SG, Wanat MJ, Gan JO, Horne EA, Hart AS, Akers CA, Parker JG, Willuhn I, Martinez V, Evans SB, Stella N, Phillips PEM (2010) Chronic microsensors for longitudinal, subsecond dopamine detection in behaving animals. Nat Methods 7:126-129.

Clark JJ, Hollon NG, Phillips PEM (2012) Pavlovian valuation systems in learning and decision making. Curr Opin Neurobiol 22: 1054-1061.

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

Cohen MX, Frank MJ (2009) Neurocomputational models of basal ganglia function in learning, memory and choice. Behav Brain Res 199:141-156.

Collins AG, Frank MJ (2014) Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev 121:337-366.

Colwill RM, Rescorla RA (1986) Associative structures in instrumental learning. In: The Psychology of Learning and Motivation: Advances in Research and Theory (Bower GH, ed), pp 55-104. Orlando, FL: Academic Press, Inc.

Daw ND (2007) Dopamine: at the intersection of reward and action. Nat Neurosci 10:1505-1507.

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199-204.

Daw ND, O'Doherty JP (2014) Multiple systems for value learning. In: Neuroeconomics: Decision Making and the Brain, 2nd Edition (Glimcher PW, Fehr E, eds), pp 393-410. London: Elsevier Academic Press.

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

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704-1711.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204-1215.

Day JJ, Jones JL, Wightman RM, Carelli RM (2010) Phasic nucleus accumbens dopamine release encodes effort- and delay-related costs. Biol Psychiatry 68:306-309.

Dayan P (2012) Instrumental vigour in punishment and reward. Eur J Neurosci 35:1152-1168.

Dayan P, Sejnowski TJ (1996) Exploration bonuses and dual control. Machine Learning 25:5-22.

de Lafuente V, Romo R (2011) Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. Proc Natl Acad Sci U S A 108:19767-19771.

Dickinson A (1985) Actions and habits: the development of behavioral autonomy. In: Philosophical Transactions of the Royal Society of London B, pp 67-78.

Dolan RJ, Dayan P (2013) Goals and habits in the brain. Neuron 80:312-325.

Dorris MC, Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron 44:365-378.

du Hoffmann J, Nicola SM (2014) Dopamine invigorates reward seeking by promoting cue-evoked excitation in the nucleus accumbens. J Neurosci 34:14349-14364.

Eblen F, Graybiel AM (1995) Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. J Neurosci 15:5999-6013.

Fiorillo CD (2013) Two dimensions of value: dopamine neurons represent reward but not aversiveness. Science 341:546-549.

Fiorillo CD, Tobler PN, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299:1898-1902.

Fiorillo CD, Yun SR, Song MR (2013a) Diversity and homogeneity in responses of midbrain dopamine neurons. J Neurosci 33:4693-4709.

Fiorillo CD, Song MR, Yun SR (2013b) Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli. J Neurosci 33:4710-4725.

Floresco SB, Tse MT, Ghods-Sharifi S (2008) Dopaminergic and glutamatergic regulation of effort- and delay-based decision making. Neuropsychopharmacology 33:1966-1979.

Frank MJ (2011) Computational models of motivated action selection in corticostriatal circuits. Curr Opin Neurobiol 21:381-386.

Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306:1940-1943.

Friedman M (1953) Essays in Positive Economics. Chicago: University of Chicago Press.

Gan JO, Walton ME, Phillips PEM (2010) Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. Nat Neurosci 13:25-27.

Garrison J, Erdeniz B, Done J (2013) Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neurosci Biobehav Rev 37:1297-1310.

Gerfen CR, Surmeier DJ (2011) Modulation of striatal projection systems by dopamine. Annu Rev Neurosci 34:441-466.

Gerfen CR, Paletzki R, Heintz N (2013) GENSAT BAC cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. Neuron 80:1368-1383.

Ghods-Sharifi S, St Onge JR, Floresco SB (2009) Fundamental contribution by the basolateral amygdala to different forms of decision making. J Neurosci 29:5251-5259.

Glimcher PW (2009) Representation of value in the primate brain. In: Basal Ganglia in Health and Disease, McGovern Institute Annual Symposium, Massachusetts Institute of Technology. URL: http://video.mit.edu/watch/representation-of-value-in-the-primate-brain-9487/.

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

Glimcher PW, Dorris MC, Bayer HM (2005) Physiological utility theory and the neuroeconomics of choice. Games Econ Behav 52:213-256.

Glimcher PW, Camerer CF, Fehr E, Poldrack RA, eds (2009) Neuroeconomics: Decision Making and the Brain, 1st Edition. London: Academic Press.

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66:585-595.

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535-574.

Graybiel AM (1995) Building action repertoires: memory and learning functions of the basal ganglia. Curr Opin Neurobiol 5:733-741.

Green DM, Swets JA (1966) Signal Detection Theory and Psychophysics. New York: Wiley.

Gremel CM, Costa RM (2013) Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat Commun 4:2264.

Groenewegen HJ, Uylings HBM (2010) Organization of prefrontal-striatal connections. In: Handbook of Basal Ganglia Structure and Function (Steiner H, Tseng KY, eds), pp 353-365. Amsterdam: Academic Press.

Haber SN, Knutson B (2010) The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology 35:4-26.

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A (2011) Transformation of stimulus value signals into motor commands during simple choice. Proc Natl Acad Sci U S A 108:18120-18125.

Hart AS, Rutledge RB, Glimcher PW, Phillips PEM (2014) Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci 34:698-704.

Heien ML, Khan AS, Ariansen JL, Cheer JF, Phillips PEM, Wassum KM, Wightman RM (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.

Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 4:267-272.

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

Hillman KL, Bilkey DK (2010) Neurons in the rat anterior cingulate cortex dynamically encode cost-benefit in a spatial decision-making task. J Neurosci 30:7705-7713.

Hollon NG, Arnold MM, Gan JO, Walton ME, Phillips PEM (2014) Dopamine-associated cached values are not sufficient as the basis for action selection. Proc Natl Acad Sci U S A 111:18357-18362.

Holroyd CB, McClure SM (2015) Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. Psychol Rev 122:54-83.

Hong S, Hikosaka O (2011) Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning. Front Behav Neurosci 5:15.

Horvitz JC (2000) Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience 96:651-656.

Houk JC, Adams JL, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds), pp 249-270: MIT Press.

Howe MW, Tierney PL, Sandberg SG, Phillips PEM, Graybiel AM (2013) Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature 500:575-579.

Huys QJ, Pizzagalli DA, Bogdan R, Dayan P (2013) Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. Biol Mood Anxiety Disord 3:12.

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

Ilango A, Kesner AJ, Keller KL, Stuber GD, Bonci A, Ikemoto S (2014b) Similar roles of substantia nigra and ventral tegmental dopamine neurons in reward and aversion. J Neurosci 34:817-822.

Ito M, Doya K (2011) Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. Curr Opin Neurobiol 21:368-373.

Izhikevich EM (2007) Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex 17:2443-2452.

Jin X, Costa RM (2010) Start/stop signals emerge in nigrostriatal circuits during sequence learning. Nature 466:457-462.

Jin X, Tecuapetla F, Costa RM (2014) Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. Nat Neurosci 17:423-430.

Joshua M, Adler A, Mitelman R, Vaadia E, Bergman H (2008) Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials. J Neurosci 28:11673-11684.

Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10:1625-1633.

Kable JW, Glimcher PW (2009) The neurobiology of decision: consensus and controversy. Neuron 63:733-745.

Kahneman D (2003) Maps of bounded rationality: psychology for behavioral economics. American Economic Review 95:1449-1475.

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263-291.

Kakade S, Dayan P (2002) Dopamine: generalization and bonuses. Neural Netw 15:549-559.

Kennerley SW, Behrens TE, Wallis JD (2011) Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nat Neurosci 14:1581-1589.

Kennerley SW, Dahmubed AF, Lara AH, Wallis JD (2009) Neurons in the frontal lobe encode the value of multiple decision variables. J Cogn Neurosci 21:1162-1178.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. Nature 455:227-231.

Keramati M, Dezfouli A, Piray P (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol 7:e1002055.

Kim H, Sul JH, Huh N, Lee D, Jung MW (2009) Role of striatum in updating values of chosen actions. J Neurosci 29:14701-14712.

Kim KM, Baratta MV, Yang A, Lee D, Boyden ES, Fiorillo CD (2012) Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PLoS One 7:e33612.

Kim S, Hwang J, Lee D (2008) Prefrontal coding of temporally discounted values during intertemporal choice. Neuron 59:161-172.

Kim YB, Huh N, Lee H, Baeg EH, Lee D, Jung MW (2007) Encoding of action history in the rat ventral striatum. J Neurophysiol 98:3548-3556.

Kishida KT, King-Casas B, Montague PR (2010) Neuroeconomic approaches to mental disorders. Neuron 67:543-554.

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

Koga K, Descalzi G, Chen T, Ko H, Lu J, Li S, Son J, Kim T, Kwak C, Huganir RL, Zhao M, Kaang B, Collingridge GL, Zhuo M (2015) Coexistence of two forms of LTP in ACC

provides a synaptic mechanism for the interactions between anxiety and chronic pain. Neuron 85:377-389.

Kreitzer AC, Malenka RC (2008) Striatal plasticity and basal ganglia circuit function. Neuron 60:543-554.

Lak A, Stauffer WR, Schultz W (2014) Dopamine prediction error responses integrate subjective value from different reward dimensions. Proc Natl Acad Sci U S A 111:2343-2348.

Lammel S, Lim BK, Ran C, Huang KW, Betley MJ, Tye KM, Deisseroth K, Malenka RC (2012) Input-specific control of reward and aversion in the ventral tegmental area. Nature 491:212-217.

Lau B, Glimcher PW (2007) Action and outcome encoding in the primate caudate nucleus. J Neurosci 27:14502-14514.

Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451-463.

Lee D (2013) Decision making: from neuroscience to psychiatry. Neuron 78:233-248.

Lee D, Seo H, Jung MW (2012) Neural basis of reinforcement learning and decision making. Annu Rev Neurosci 35:287-308.

Lee SW, Shimojo S, O'Doherty JP (2014) Neural computations underlying arbitration between model-based and model-free learning. Neuron 81:687-699.

Lima SQ, Hromádka T, Znamenskiy P, Zador AM (2009) PINP: a new method of tagging neuronal populations for identification during in vivo electrophysiological recording. PLoS One 4:e6099.

Lodge DJ (2011) The medial prefrontal and orbitofrontal cortices differentially regulate dopamine system function. Neuropsychopharmacology 36:1227-1236.

Louie K, Glimcher PW (2010) Separating value from choice: delay discounting activity in the lateral intraparietal area. J Neurosci 30:5498-5507.

Mackintosh NJ, ed (1994) Animal Learning and Cognition, 2nd Edition. San Diego, CA: Academic Press, Inc.

Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154-162.

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

McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. Science 306:503-507.

McClure SM, Ericson KM, Laibson DI, Loewenstein G, Cohen JD (2007) Time discounting for primary rewards. J Neurosci 27:5796-5804.

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 31:2700-2705.

McHugh SB, Barkus C, Huber A, Capitão L, Lima J, Lowry JP, Bannerman DM (2014) Aversive prediction error signals in the amygdala. J Neurosci 34:9024-9033.

Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50:381-425.

Mirenowicz J, Schultz W (1996) Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. Nature 379:449-451.

Mogenson GJ, Jones DL, Yim CY (1980) From motivation to action: functional interface between the limbic system and the motor system. Prog Neurobiol 14:69-97.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936-1947.

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16:72-80.

Morita K (2014) Differential cortical activation of the striatal direct and indirect pathway cells: reconciling the anatomical and optogenetic results by using a computational method. J Neurophysiol 112:120-146.

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

Morita K, Morishima M, Sakai K, Kawaguchi Y (2013) Dopaminergic control of motivation and reinforcement learning: a closed-circuit account for reward-oriented behavior. J Neurosci 33:8866-8890.

Morris G, Schmidt R, Bergman H (2010) Striatal action-learning based on dopamine concentration. Exp Brain Res 200:307-317.

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

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

Morrison SE, Nicola SM (2014) Neurons in the nucleus accumbens promote selection bias for nearer objects. J Neurosci 34:14147-14162.

Nasrallah NA, Clark JJ, Collins AL, Akers CA, Phillips PEP, Bernstein IL (2011) Risk preference following adolescent alcohol use is associated with corrupted encoding of costs but not rewards by mesolimbic dopamine. Proc Natl Acad Sci U S A 108:5466-5471.

Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15:1-25.

Nicola SM (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, Daw ND, Dayan P (2006) Choice values. Nat Neurosci 9:987-988.

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

Nomoto K, Schultz W, Watanabe T, Sakagami M (2010) Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J Neurosci 30:10692-10702.

Oleson EB, Gentry RN, Chioma VC, Cheer JF (2012) Subsecond dopamine release in the nucleus accumbens predicts conditioned punishment and its successful avoidance. J Neurosci 32:14804-14808.

Ostlund SB, Maidment NT (2012) Dopamine receptor blockade attenuates the general incentive motivational effects of noncontingently delivered rewards and reward-paired cues without affecting their ability to bias action selection. Neuropsychopharmacology 37:508-519.

Packard MG (2009) Exhumed from thought: basal ganglia and response learning in the plus-maze. Behav Brain Res 199:24-31.

Padoa-Schioppa C (2011) Neurobiology of economic choice: a good-based model. Annu Rev Neurosci 34:333-359.

Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. Nature 441:223-226.

Palminteri S, Justo D, Jauffret C, Pavlicek B, Dauta A, Delmaire C, Czernecki V, Karachi C, Capelle L, Durr A, Pessiglione M (2012) Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. Neuron 76:998-1009.

Pasquereau B, Turner RS (2013) Limited Encoding of Effort by Dopamine Neurons in a Cost-Benefit Trade-off Task. J Neurosci 33:8288-8300.

Pavlov IP (1927) Conditioned Reflexes. London: Oxford University Press.

Pawlak V, Kerr JN (2008) Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. J Neurosci 28:2435-2446.

Pawlak V, Wickens JR, Kirkwood A, Kerr JN (2010) Timing is not Everything: Neuromodulation Opens the STDP Gate. Front Synaptic Neurosci 2:146.

Paxinos G, Watson C (2005) The rat brain in stereotaxic coordinates, 5th Edition. London: Elsevier Academic Press.

Phillips PEM, Walton ME, Jhou TC (2007) Calculating utility: preclinical evidence for cost-benefit analysis by mesolimbic dopamine. Psychopharmacology (Berl) 191:483-495.

Phillips PEM, Stuber GD, Heien ML, Wightman RM, Carelli RM (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422:614-618.

Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. Nature 400:233-238.

Poldrack RA, Packard MG (2003) Competition among multiple memory systems: converging evidence from animal and human brain studies. Neuropsychologia 41:245-251.

Prévost C, Pessiglione M, Météreau E, Cléry-Melin ML, Dreher JC (2010) Separate valuation subsystems for delay and effort decision costs. J Neurosci 30:14080-14090.

Rangel A, Hare T (2010) Neural computations associated with goal-directed choice. Curr Opin Neurobiol 20:262-270.

Rangel A, Clithero JA (2014) The computation of stimulus values in simple choice. In: Neuroeconomics: Decision Making and the Brain, 2nd Edition (Glimcher PW, Fehr E, eds), pp 125-148. London: Elsevier Academic Press.

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545-556.

Ravel S, Richmond BJ (2006) Dopamine neuronal responses in monkeys performing visually cued reward schedules. Eur J Neurosci 24:277-290.

Redgrave P, Prescott TJ, Gurney K (1999) The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 89:1009-1023.

Redish AD (2004) Addiction as a computational process gone awry. Science 306:1944-1947.

Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci 31:415-437; discussion 437-487.

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In: Classical Conditioning II:

Current Research and Theory (Black AH, Prokasy WF, eds), pp 64-99. New York: Appleton-Century-Crofts.

Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 413:67-70.

Robbins TW, Everitt BJ (2007) A role for mesencephalic dopamine in activation: commentary on Berridge (2006). Psychopharmacology (Berl) 191:433-437.

Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10:1615-1624.

Roesch MR, Singh T, Brown PL, Mullins SE, Schoenbaum G (2009) Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. J Neurosci 29:13365-13376.

Rossi MA, Sukharnikova T, Hayrapetyan VY, Yang L, Yin HH (2013) Operant self-stimulation of dopamine neurons in the substantia nigra. PLoS One 8:e65799.

Rudebeck PH, Walton ME, Smyth AN, Bannerman DM, Rushworth MF (2006) Separate neural pathways process different decision costs. Nat Neurosci 9:1161-1168.

Rummery GA, Niranjan M (1994) On-line Q-learning using connectionist systems. In: Technical Report CUED/F-INENG/TR 166. Engineering Department, Cambridge University.

Saddoris MP, Sugam JA, Stuber GD, Witten IB, Deisseroth K, Carelli RM (2014) Mesolimbic dopamine dynamically tracks, and is causally linked to, discrete aspects of value-based decision making. Biol Psychiatry.

Salamone JD, Correa M, Farrar A, Mingote SM (2007) Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. Psychopharmacology (Berl) 191:461-482.

Salamone JD, Correa M, Farrar AM, Nunes EJ, Pardo M (2009) Dopamine, behavioral economics, and effort. Front Behav Neurosci 3:13.

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

Samuelson PA (1937) A note on the measurement of utility. The Review of Economic Studies 4:155-161.

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

Schultz W (2006) Behavioral theories and the neurophysiology of reward. Annu Rev Psychol 57:87-115.

Schultz W (2013) Updating dopamine reward signals. Curr Opin Neurobiol 23:229-238.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593-1599.

Seo M, Lee E, Averbeck BB (2012) Action selection and action value in frontal-striatal circuits. Neuron 74:947-960.

Sesack SR, Grace AA (2010) Cortico-basal ganglia reward network: microcircuitry. Neuropsychopharmacology 35:27-47.

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

Skinner BF (1938) The Behavior of Organisms: An Experimental Analysis. In. New York: Appleton-Century.

Skvortsova V, Palminteri S, Pessiglione M (2014) Learning to minimize efforts versus maximizing rewards: computational principles and neural correlates. J Neurosci 34:15621-15630.

Stauffer WR, Lak A, Schultz W (2014) Dopamine reward prediction error responses reflect marginal utility. Curr Biol 24:2491-2500.

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16:966-973.

Steinberg EE, Boivin JR, Saunders BT, Witten IB, Deisseroth K, Janak PH (2014) Positive reinforcement mediated by midbrain dopamine neurons requires D1 and D2 receptor activation in the nucleus accumbens. PLoS One 9:e94771.

Stephenson-Jones M, Kardamakis AA, Robertson B, Grillner S (2013) Independent circuits in the basal ganglia for the evaluation and selection of actions. Proc Natl Acad Sci U S A 110:E3670-3679.

Stopper CM, Tse MT, Montes DR, Wiedman CR, Floresco SB (2014) Overriding phasic dopamine signals redirects action selection during risk/reward decision making. Neuron 84:177-189.

Stott JJ, Redish AD (2014) A functional difference in information processing between orbitofrontal cortex and ventral striatum during decision-making behaviour. Philos Trans R Soc Lond B Biol Sci 369.

Strait CE, Blanchard TC, Hayden BY (2014) Reward value comparison via mutual inhibition in ventromedial prefrontal cortex. Neuron 82:1357-1366.

Sugam JA, Day JJ, Wightman RM, Carelli RM (2012) Phasic nucleus accumbens dopamine encodes risk-based decision-making behavior. Biol Psychiatry 71:199-205.

Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304:1782-1787.

Sugrue LP, Corrado GS, Newsome WT (2005) Choosing the greater of two goods: neural currencies for valuation and decision making. Nat Rev Neurosci 6:363-375.

Sul JH, Jo S, Lee D, Jung MW (2011) Role of rodent secondary motor cortex in value-based action selection. Nat Neurosci 14:1202-1208.

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

Sutton RS (1988) Learning to predict by the methods of temporal differences. Machine Learning 3:9-44.

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.

Sutton RS, Precup D, Singh S (1999) Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artificial Intelligence 112:181-211.

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

Takahashi YK, Roesch MR, Wilson RC, Toreson K, O'Donnell P, Niv Y, Schoenbaum G (2011) Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat Neurosci 14:1590-1597.

Takahashi YK, Roesch MR, Stalnaker TA, Haney RZ, Calu DJ, Taylor AR, Burke KA, Schoenbaum G (2009) The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron 62:269-280.

Thorndike EL (1911) Animal Intelligence: Experimental Studies. New York: Macmillan.

Toates F (1998) The interaction of cognitive and stimulus-response processes in the control of behaviour. Neurosci Biobehav Rev 22:59-83.

Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642-1645.

Tolman EC (1949) There is more than one kind of learning. Psychol Rev 56:144-155.

Tsai HC, Zhang F, Adamantidis A, Stuber GD, Bonci A, de Lecea L, Deisseroth K (2009) Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science 324:1080-1084.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453-458.

Ungless MA, Magill PJ, Bolam JP (2004) Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. Science 303:2040-2042.

van der Meer MA, Redish AD (2009) Covert expectation-of-reward in rat ventral striatum at decision points. Front Integr Neurosci 3:1.

van der Meer MA, Redish AD (2010) Expectancies in decision making, reinforcement learning, and ventral striatum. Front Neurosci 4:6.

von Neumann J, Morgenstern O (1944) The Theory of Games and Economic Behavior. Princeton, NJ: Princeton University Press.

Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43-48.

Wall NR, Wickersham IR, Cetin A, De La Parra M, Callaway EM (2010) Monosynaptic circuit tracing in vivo through Cre-dependent targeting and complementation of modified rabies virus. Proc Natl Acad Sci U S A 107:21848-21853.

Walton ME, Bannerman DM, Alterescu K, Rushworth MF (2003) Functional specialization within medial frontal cortex of the anterior cingulate for evaluating effort-related decisions. J Neurosci 23:6475-6479.

Wanat MJ, Kuhnen CM, Phillips PEM (2010) Delays conferred by escalating costs modulate dopamine release to rewards but not their predictors. J Neurosci 30:12020-12027.

Wang DV, Tsien JZ (2011) Convergent processing of both positive and negative motivational signals by the VTA dopamine neuronal populations. PLoS One 6:e17047.

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

Watkins CJCH (1989) Learning from delayed rewards. PhD. Thesis, Cambridge University.

Wickens JR, Begg AJ, Arbuthnott GW (1996) Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. Neuroscience 70:1-5.

Wickersham IR, Lyon DC, Barnard RJ, Mori T, Finke S, Conzelmann KK, Young JA, Callaway EM (2007) Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. Neuron 53:639-647.

Witten IB, Steinberg EE, Lee SY, Davidson TJ, Zalocusky KA, Brodsky M, Yizhar O, Cho SL, Gong S, Ramakrishnan C, Stuber GD, Tye KM, Janak PH, Deisseroth K (2011)

Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron 72:721-733.

Wunderlich K, Rangel A, O'Doherty JP (2010) Economic choices can be made using only stimulus values. Proc Natl Acad Sci U S A 107:15005-15010.

Yin HH, Ostlund SB, Balleine BW (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. Eur J Neurosci 28:1437-1448.

Zhang L, Doyon WM, Clark JJ, Phillips PEM, Dani JA (2009) Controls of tonic and phasic dopamine transmission in the dorsal and ventral striatum. Mol Pharmacol 76:396-404.