Dopamine signaling across striatal subregions during acquisition of instrumental associations

Tobias Wolfgang Bernklau

Graduate School of
Systemic Neurosciences
LMU Munich

Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

December 2021

First reviewer and supervisor:	Prof. Dr. Simon Jacob Department of Neurosurgery Technical University of Munich
Second reviewer:	Prof. Dr. Anton Sirota
Third reviewer:	Dr. Sevil Duvarci
Date of submission:	December 17, 2021
Date of defense:	May 25, 2022

Table of contents

1. Abstract

Associative learning and flexible decision-making form the basis of intelligent behavior. As one of the major neuromodulators in the brain, dopamine plays a key role as a teaching signal in learning and decision-making. Dopamine is involved in a variety of brain functions with several clinical implications and has been studied extensively in the context of reward prediction error coding. In animal models, which offer a large array of neuroscientific tools, most research examines the role of dopamine in Pavlovian or simple instrumental tasks, while more complex learning paradigms are less abundant. Moreover, most studies of instrumental behavior examine consolidated performance in well-trained animals. How dopamine signals in different dopaminergic projection targets evolve during the acquisition of novel instrumental associations in naïve animals and how predictive and reinforcing aspects of dopamine reward prediction errors contribute to instrumental learning remain elusive. To address these points, I developed a decision-making task for mice and performed direct fluorescent imaging of dopamine using fiber photometry in several subregions of the striatum, while mice acquired instrumental associations to learn the task. In this dissertation, after a short review of the relevant literature, I present behavioral and neural data in two main analyses.

First, I present a detailed analysis of the mouse behavior during learning on multiple timescales down to sub-trial resolution. In an auditory decision-making task with rule switches, mice acquired different rule-based associations between auditory instruction cues and instrumental responses in order to obtain rewards. Choice models revealed that mice used different strategies to perform the task, depending on the learning state. During learning, but not in the fully-trained state, mice showed a tendency to repeat previous choices, but no other trial history-related biases. When animals were fully trained, they followed only the instruction cue to guide their choice. A trial-based choice model with fluctuating weights for response bias and stimulus instructions best explained the animals' choices.

Second, I present analyses of the time course of dopamine fluctuations in the ventral, dorsomedial, and dorsolateral striatum of mice during behavior. Dopamine signals were robustly modulated by different task events and showed signatures of reward prediction errors in all striatal subregions. The dopamine reward response was inversely scaled to task performance and reset after rule switches, while the dopamine cue response was not analogously scaled. Dopamine cue and reward responses were modulated by behavioral preference for particular conditions, revealing a signature in accordance with reward prediction errors. Strikingly, in the ventral striatum this signature was corrupted during beginner sessions after rule switches. The reward prediction error signature partially lagged behind the behavioral signature when animals showed no behavioral signs of learning and was restored when learning was apparent, suggesting a mechanism of dopamine signaling during the acquisition of associations. In particular, the dopamine cue response was intact, while the dopamine reward response lagged behind the behavioral preference, which suggested that the reinforcing aspect of dopamine reward prediction errors correlated with learning, while the predictive aspect was independent of learning. The present study furthers the understanding how predictive and reinforcing aspects or dopamine reward prediction errors contribute to the acquisition of instrumental associations.

2. Introduction

2.1 Learning and decision-making

Intelligent organisms like humans and other species thrive by making adaptive decisions to achieve their goals. Flexible decision-making is enabled by a large set of cognitive functions that allow organisms to interact with their environment in a goal-oriented way. This interaction requires an organism to process external stimuli in order to guide its own actions, to select appropriate actions, and to evaluate the outcomes of its actions in order to learn for the future. Associations between stimulus, action, and outcome – what Skinner (1953) called "three-term contingency" – are learned and constantly updated while decisions are made and evaluated. Learning and decision making are therefore inevitably intertwined and influence each other. Decisions are made on the basis of learned associations and associations are learned and updated on the basis of the outcomes of decisions. Thus, associative learning processes have a key role in intelligent behavior and the underlying mechanisms are a central research topic in psychology and neuroscience.

2.1.1 Stimulus – action – outcome

Associative learning is classically divided into Pavlovian learning and instrumental learning. Pavlovian learning (also well-known as classical conditioning) describes the formation of an association between an initially neutral stimulus (e.g., a sound) with an unconditioned stimulus (e.g., food) that elicits an innate unconditioned response (e.g., salivating). Pavlov (1927) first discovered that after repeatedly presenting a neutral stimulus and an unconditioned stimulus simultaneously or in close succession, the neutral stimulus becomes a conditioned stimulus and is able to elicit the response (now called conditioned response) even when the conditioned stimulus is presented without the unconditioned stimulus, which indicates that an association has been formed. Pavlovian learning therefore describes how a stimulus is associated with an outcome (which can be appetitive or aversive), while there is no voluntary action involved, but only innate responses.

In contrast to Pavlovian learning, in instrumental learning (usually synonymously termed operant learning or operant conditioning), the learned association involves a voluntary action in addition to the stimulus and the outcome. Instrumental learning was pioneered by Thorndike (1898), who described that cats learn to escape from a box by trial and error. The theory was later formulated by Skinner (1938), who trained rats to press a lever for food rewards in an operant chamber (also widely known as "Skinner box"). The principle of instrumental learning is that an action in response to a stimulus is more likely to be repeated in the future, if it is reinforced by an appetitive outcome, or less likely to be repeated, if it is paired with an aversive outcome. The crucial difference to Pavlovian learning is that in instrumental learning the outcome is contingent on the instrumental action, while in Pavlovian learning the outcome simply follows the stimulus. It has to be noted that while the distinction between Pavlovian and instrumental learning is common and useful, it can also be blurred, since Pavlovian learning only becomes apparent through behavior, which in some cases can be easily separated from the types of actions involved in instrumental learning (e.g., in the case of salivating), but in other cases the distinction may be more difficult (e.g., in the case of approach behavior).

Pavlovian learning can also influence instrumental learning via so-called “Pavlovian to instrumental transfer”, for example, when a previously classically conditioned stimulus is more readily associated during instrumental learning compared to a novel stimulus (Dickinson & Balleine, 1994; Rescorla & Solomon, 1967; Talmi et al., 2008).

Instrumental behavior can be further divided along different decision-making strategies, which are usually described according to their emphasis on one of the associations between the three terms of Skinner’s “three-term-contingency”: stimulus, action (or response) and outcome (Fig. 1). While Pavlovian learning only involves an association between the stimulus and the outcome, different forms of instrumental behavior differently involve an association with the action. The main distinction here is between goal-directed and habitual behavior. Both in rodents and humans, goal-directed behavior is assumed to rely on action-outcome associations, while habitual behavior is assumed to rely on stimulus-response (or stimulus-action) associations (Balleine & O'Doherty, 2010). The terms “action” and “response” are usually used synonymously, but depending on the type of behavior, one or the other is more common (typically “stimulus-response” and “action-outcome” are used). Similarly, the “stimulus” is often termed “cue”, especially in tasks in which the stimulus acts as an instruction cue.

The difference between goal-directed and habitual behavior was shown experimentally by Adams and Dickinson (1981) in rats performing lever presses to obtain sucrose rewards. The distinction between the two behaviors was tested by an outcome devaluation test, which revealed that if the outcome was devalued by replacing sucrose with illness-inducing lithium chloride, animals responded with fewer lever presses, also in test trials in the absence of feedback. This suggested that the animals were sensitive to changes in the action-outcome association and the behavior was therefore categorized as goal-directed. Interestingly, with increased training, the rats became insensitive to outcome devaluation and continued pressing the lever, which suggested that they now relied on a habitual strategy based on a stimulus-response association, since the outcome no longer had an impact on the response (Adams, 1982). Another test of goal-directed behavior is to degrade the contingency between actions and rewards by giving rewards without preceding actions. If the behavior is goal-directed, the actions should be reduced, if it is habitual, it should be independent of changes in the outcome contingency (Balleine & Dickinson, 1998). While early behaviorists wanted to explain instrumental behavior exclusively as either stimulus-response (Hull, 1943; Thorndike, 1898) or action-outcome learning (Tolman, 1948), it is now widely accepted that the two processes are not alternatives but occur depending on certain conditions. Goal-directed behavior is promoted by short amounts of training and unpredictability, while extensive training, predictability, and stress promote habitual behavior (Redgrave et al., 2010). The two processes can also quickly alternate or work in parallel (Balleine & O'Doherty, 2010) and form the basis for the broad distinction of automatic (habitual) in contrast to controlled or deliberative (goal-directed) systems that govern decisions, popularly simplified as “thinking fast and slow” (Kahneman, 2011).

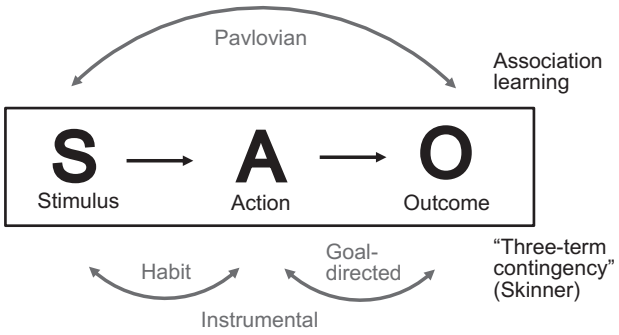


Fig. 1) Association learning
Different associations (indicated by gray arrows) between stimuli, actions, and outcomes are assumed to be involved in different forms of learning and decision-making.

2.1.2 Reinforcement learning

The work of animal behaviorists and the principles of learning theory inspired researchers to formulate computational models of learning, which led to the foundation of the field of reinforcement learning. In early models, Pavlovian learning was mathematically formalized as an iterative process of updating the strength of an association after every repetition of stimulus pairing by adding the difference between the expected reward and the actual reward, that is, a reward prediction error (Bush & Mosteller, 1951; Rescorla & Wagner, 1972). Building on this basic algorithm, computer scientists Sutton and Barto (1981) started working on this problem and developed the temporal difference reinforcement learning model (Sutton, 1988; Sutton & Barto, 1998), which predicts future rewards at discrete timepoints with updates using the reward prediction error. Compared to earlier versions of the model, the temporal difference model predicts expected value instead of association strength and does so at every moment in time instead of every repetition. The reward prediction or value at a given moment depends on the expectation of future rewards. Predicted rewards that lie further in the future are discounted compared to closer rewards. If an unexpected reward occurs at a given moment, there is a positive reward prediction error and the value of that moment is increased. The credit of the reward is not only assigned to the current moment, but also to a certain number of previous moments. If unpredicted rewards occur repeatedly after certain cues (as in Pavlovian learning paradigms), the cues become predictive of the future rewards and themselves elicit reward prediction errors. If a reward is fully predicted by a preceding cue, it no longer elicits a reward prediction error. Thus, the reward prediction error acts as a teaching signal during learning. If the prediction matches the actual reward there is no error and no learning takes place. If the prediction error is positive or negative, the value is increased or decreased, respectively.

The temporal difference reinforcement learning model can be applied not only to Pavlovian learning but also to instrumental learning by assigning values to states and actions and updating them according to reward prediction errors. In a reinforcement learning model, an agent learns to maximize rewards by taking the action with the highest estimated value in a given state. Estimating the values of actions and deriving an optimal policy of actions to choose are the key elements of reinforcement learning algorithms. The value of a state is estimated not only from the immediately imminent reward, but from the sum of all expected future rewards. A reinforcement learning model therefore also

includes functions that describe the transition between states depending on the actions that are chosen. Reinforcement learning models can vary in the degree to which these transition functions, or models of the environment, are available to the agent. In “model-based” reinforcement learning algorithms, the agent has access to the model of the environment and is able to use it for planning actions by taking into account not only the current state and available actions, but also future states and actions. In “model-free” algorithms, such as the original temporal-difference reinforcement learning model, the transition functions are not known to the agent, actions are taken based on values that represent a mixture of the reward and transition structure, and values are learned by trial-and-error without an explicit model of the environment (Sutton & Barto, 2018).

By their definition, model-based and model-free learning algorithms are intuitively related to goal-directed and habitual decision-making. The two concepts have been brought together in computational models (Daw et al., 2005) and behavioral tasks (Daw et al., 2011), linking goal-directed decisions to model-based action selection, and habitual decisions to model-free action selection. However, model-based and model-free processes are learning algorithms and goal-directed and habitual processes are decision-making strategies. It is therefore questionable to use the two concepts synonymously, albeit this is often the case (Decker et al., 2016; Drummond & Niv, 2020). While there may be a strong correspondence between model-based learning and goal-directed decision-making, some authors argue that habitual decision-making is value-free and therefore neither model-based nor model-free (Collins & Cockburn, 2020; Miller et al., 2018).

Building on the groundwork of psychologists and animal experimentalists and their descriptions of biological learning processes, reinforcement learning grew as an own field, incorporating parallel developments of control theory and dynamic programming in order to solve value functions. Advances in reinforcement learning as a major branch of artificial intelligence (together with supervised and unsupervised machine learning methods) have produced impactful applications of algorithms including, for example, human-level video game play (Mnih et al., 2015), superhuman performance in the game of Go (Silver et al., 2017), and tackling difficult robotics problems (Akkaya et al., 2019).

Reinforcement learning is of great interest not only to the fields of psychology and computer science, but also attracted the attention of neuroscientists, when Schultz and colleagues (1997) reinterpreted seminal experiments and showed that the activity of dopamine neurons in the monkey brain closely resembled reward prediction errors in temporal difference learning models. This finding led to the dopamine reward prediction error hypothesis that strongly shaped today’s understanding of dopamine function, albeit not without controversy, and will be described in more detail below after a general introduction to the dopamine system. The fields of neuroscience and reinforcement learning have since been engaged in a “virtuous circle” (Hassabis et al., 2017) of mutual inspiration (Dabney et al., 2020; Neftci & Averbeck, 2019).

2.2 The dopamine system

Dopamine is one of the major neuromodulators in the brain. It is involved in a large variety of brain functions with high clinical relevance.

2.2.1 Anatomy and physiology of dopamine neurons

Dopamine is released from subcortical neurons that are typically identified by the synthesizing enzyme tyrosine hydroxylase (Björklund & Dunnett, 2007) or the dopamine transporter (Lammel et al., 2015). While the anatomy of the dopamine system is largely conserved across mammals (Björklund & Dunnett, 2007) and very similar even in birds (Durstewitz et al., 1999), there are also fine differences in anatomy and function, for example, between primates and rodents. In mammals, dopaminergic neurons have their origin mainly in midbrain nuclei in the regions of the ventral tegmental area (VTA) and substantia nigra (SN) and send projections to the forebrain, where the major targets are the striatum and the frontal cortex (Fig. 2). These projections are classically divided into mesostriatal (or, more traditionally, nigrostriatal), mesolimbic and mesocortical pathways. Initially, it was assumed that the pathways could be separated not only by the targets, but also by the origins of the projections (i.e., VTA or SN). This is why the mesostriatal pathway was originally termed nigrostriatal pathway, suggesting that all striatal projections originated from the SN. There is now evidence that the neurons projecting to the striatum are distributed across VTA and SN, but the three pathways are still deemed anatomically and functionally distinct, and dopaminergic neurons from the different pathways rarely send collateral axons to other targets (Björklund & Dunnett, 2007). In the classical view, the main target of the nigrostriatal pathway is the caudate-putamen (or dorsal striatum); the main target of the mesolimbic pathway is the nucleus accumbens (or ventral striatum), but also the amygdala, hippocampus, septum, and olfactory tubercle; and the main target of the mesocortical pathway is the prefrontal cortex. Since the nucleus accumbens in the ventral striatum can be counted either as part of the striatum or as part of the limbic system, the more inclusive term mesostriatal pathway can be used to group all striatal projections. In primates, the striatal targets include the nucleus accumbens (or ventral striatum), the caudate, and the putamen. In rodents, the striatal targets include the nucleus accumbens (or ventral striatum), the dorsomedial striatum (DMS), dorsolateral (DLS) striatum, and the tail of the striatum (posterior part of the dorsal striatum). The rodent DMS can be considered to be the homolog of the primate caudate, while the rodent DLS can be considered as the homolog of the primate putamen, but there are also differences in connectivity that challenge this view (Balsters et al., 2020). Differences between rodents and primates can also be identified in the mesocortical pathway. Dopaminergic cortical projections are much more abundant in primates compared to rodents and target both medial and lateral prefrontal cortex. In rodents dopaminergic cortical projections mainly target the medial prefrontal cortex, anterior cingulate cortex, and perirhinal cortex, since there is no rodent homolog of the granular lateral prefrontal cortex of primates (Laubach et al., 2018). In primates, but not in rodents, the mesocortical pathway can be divided into a medial and lateral system, both on the basis of anatomy and function (Matsumoto & Takada, 2013; Ranganath & Jacob, 2016). Midbrain dopaminergic neurons receive direct inputs from a variety of regions, including feedback connections from the striatum and cortex (Watabe-Uchida et al., 2012). For example, the medial prefrontal cortex

makes monosynaptic connections to midbrain dopamine neurons projecting to the ventral striatum (Beier et al., 2015).

Dopamine neurons have been shown to fire action potentials in two different physiological modes. Both in rodents (Grace & Bunney, 1984; Hyland et al., 2002) and in primates (Bayer et al., 2007) dopamine neurons have relatively low baseline firing rates (tonic activity) and fire bursts of action potentials (phasic activity) in response to task events.

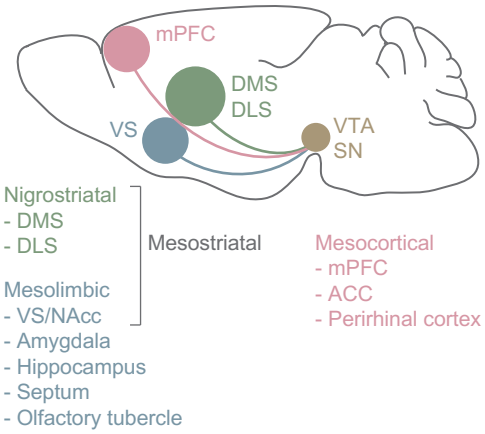


Fig. 2) Schematic of the rodent dopamine system
Sagittal section of the mouse brain indicating the most relevant projections of midbrain dopamine neurons; data from Björklund and Dunnett (2007); brain atlas schematic adapted from Paxinos and Franklin (2001); DMS = dorsomedial striatum, DLS = dorsolateral striatum, VS = ventral striatum, NAcc = nucleus accumbens, mPFC = medial prefrontal cortex, ACC = anterior cingulate cortex, VTA = ventral tegmental area, SN = substantia nigra.

2.2.2 Measuring dopaminergic activity

In order to elucidate the functions of dopaminergic signaling, dopamine neuron activity needs to be examined in vivo. The in vivo physiology and function of dopamine neurons was first examined with electrophysiology in animal models, mainly nonhuman primates, rats, and mice. Using electrodes inserted in the brain, the electrical activity that arises when neurons fire action potentials can be monitored either intra-cellularly or extra-cellularly. With this technique the physiological properties of neurons could be characterized and putative dopamine neurons could be identified by their firing properties and shape of their action potential waveforms (Matsumoto & Hikosaka, 2009), although the reliability of this method has limitations (Margolis et al., 2006). Overcoming the shortcomings of electrophysiology, technological advances in the development of neuroscientific tools have recently opened up many possibilities for the measurement and manipulation of dopaminergic activity.

The development of genetic tools and fluorescent imaging techniques in mice (and to some extent also in rats and nonhuman primates) has enabled researchers to identify dopamine neurons not only by the shape of their action potentials, but also based on molecular markers such as tyrosine hydroxylase or the dopamine transporter. Genetically encoded fluorescent calcium indicators (Chen et al., 2013) that signal changes in intracellular calcium as a proxy of neuronal activity could now be expressed exclusively in dopaminergic neurons, making it possible to monitor the activity of dopaminergic neurons in vivo, for example using two-photon imaging, one-photon endoscopic microscopes, or fiber photometry. These methods also allow the monitoring of axonal activity, which

makes it possible to measure not only cell type-specific, but also projection-specific activity. Optogenetic tools (Boyden et al., 2005) created the possibility to manipulate cell type-specific neuronal activity via light stimulation of genetically encoded light-sensitive opsins that cause a cell to increase or decrease its activity. This method not only enabled the manipulation of dopaminergic activity in vivo in combination with behavior, but also aided the identification of dopaminergic cells in electrophysiology recordings using optogenetics to “tag” neurons by examining their electrophysiological responses to optical stimulation, both in rodents (Cohen et al., 2012) and in primates (Stauffer et al., 2016). While all these methods enable the precise measurement of neuronal activity of dopamine neurons and specific projections, they cannot provide information on the magnitude and time course of dopamine release.

In humans, dopamine levels can be measured with molecular imaging techniques like positron emission tomography in combination with radiotracers (Egerton et al., 2009; Liu et al., 2019), a technique which has low temporal resolution on the order of minutes and coarse spatial resolution compared to methods available in animal models. In animal models, traditional techniques for the direct measurement of dopamine have produced useful insights, but have also long suffered from coarse spatial or temporal resolution, or low molecular specificity. Using microdialysis (Tidey & Miczek, 1996), dopamine can be measured specifically, but only at a temporal resolution on the order of minutes. With electrochemical methods such as fast-scan cyclic voltammetry (Robinson et al., 2003), sub-second temporal resolution can be achieved, but it is difficult to separate dopamine signals from signals of other catecholamines, such as norepinephrine, due to their similar structure and oxidization profile. Other, more recent methods based on injected cells (Muller et al., 2014) or marker genes (Lee et al., 2017) offer high molecular specificity but do not offer temporal resolution in the sub-second range.

Very recently, two similar genetically encoded fluorescent protein-based sensors for the direct measurement of dopamine have been developed, dLight (Patriarchi et al., 2018) and GRAB-DA (Sun et al., 2018). Both sensors use the same principle of combining naturally occurring dopamine receptors with an inserted fluorescent protein that changes its fluorescence in response to a conformational change of the receptor upon binding of dopamine. These sensors make it possible to measure dopamine release specifically and directly in the target region at sub-second temporal resolution and therefore enable unprecedented investigations of the precise time course of dopaminergic signaling (de Jong et al., 2019; Mohebi et al., 2019; Zolin et al., 2021). For example, with the help of the fluorescent dopamine sensor dLight it was revealed that dopamine release in striatal target regions can be modulated independently from somatic activity in the origin region (Mohebi et al., 2019). Fluorescent dopamine sensors allow direct measurements of dopamine release in the target region, that is, in the region where dopamine has its effect as a neuromodulator. This helps to reveal functional aspects that may not be equally present in the activity of dopamine neuron cell bodies in the midbrain. Thus, direct (fluorescent) imaging of dopamine transients furthers the understanding of specialized dopaminergic circuits and their functions.

2.2.3 Variety of dopamine functions (a short history)

Multiple discoveries starting in the 1950s and 1960s have demonstrated very early on that dopamine is involved in a variety of brain functions and neuropsychiatric diseases, including movement, psychosis and other psychiatric symptoms, cognitive functions, motivation, and learning. The main discoveries are presented here in loose chronological order. Dopamine was first shown to be a neurotransmitter in the brain in the 1950s (Montagu, 1957). At the same time, Carlsson and colleagues (1957) showed that the monoamine precursor 3,4-dihydroxyphenylalanine could reverse reserpine-induced akinesia in rabbits, suggesting the involvement of a monoamine (serotonin or dopamine) in this movement disturbance. These findings paved the way for pioneering therapeutic experiments (Birkmayer & Hornykiewicz, 1961) and effective treatment of akinetic symptoms in Parkinson's disease (Cotzias et al., 1967) with the dopamine precursor L-dopa, which remains a common therapy today. These findings, together with experiments showing that pharmacological lesions of the nigrostriatal dopamine pathway led to severe impairments of motor function in rats (Ungerstedt, 1971), demonstrated the role of dopamine in movement.

Another dopaminergic dysfunction was described already in the 1960s, when Van Rossum (1966) suggested that dopamine was overactive in patients with schizophrenia and proposed dopamine receptor blocking as a mechanism for the successful treatment of psychosis with neuroleptic drugs. Neuroleptics had been discovered already in the 1950s through clinical observations, but the neurochemical mechanisms had been unknown. The dopamine hypothesis of schizophrenia was later refined (Davis et al., 1991) as a combination of excess dopamine in the striatum, which was thought to be responsible for the positive symptoms of schizophrenia (i.e., psychosis), and a dopamine deficit in the frontal cortex, which was thought to be responsible for the negative symptoms of schizophrenia (i.e., cognitive impairments). Later imaging studies have supported the original hypothesis, which remains a key part of today's neurobiological understanding of schizophrenia (Howes & Kapur, 2009; Weinstein et al., 2017). Additional psychiatric diseases that are assumed to involve dopaminergic mechanisms include attention-deficit hyperactivity disorder (Swanson et al., 2007; Volkow et al., 2009), addiction (Nutt et al., 2015; Volkow & Morales, 2015), and depression (Belujon & Grace, 2017; Russo & Nestler, 2013).

One aspect that most psychiatric disorders have in common is that they are accompanied by different forms of cognitive impairments. Dopaminergic dysfunction is hypothesized to be involved in the mechanisms of many of these impairments (Millan et al., 2012). The role of dopamine in cognition has been researched in animal models since the 1970s, mainly in non-human primates due to their relatively strong cognitive abilities compared to other animal models. Brozoski and colleagues (1979) showed that depletion of dopamine in the prefrontal cortex led to short-term memory impairment. Blocking dopamine receptors in the prefrontal cortex was shown to impair associative learning in monkeys performing an instrumental learning task (Puig & Miller, 2012) and to impair attentional set-shifting in rats (Floresco et al., 2006). Moreover, dopamine in the prefrontal cortex was shown to modulate visual attention (Noudoost & Moore, 2011) as well as signal-to-noise ratio (Jacob et al., 2013) and rule-coding strength (Ott et al., 2014) of prefrontal cortex neurons.

The role of dopamine in reward and motivation also goes back to the 1950s. Before the first description of dopamine in the brain, Olds and Milner (1954) showed that rats performed repeated self-stimulation via lever presses triggering electrical pulses on electrodes in different areas in the brain, including the medial forebrain bundle. These experiments suggested that the self-stimulation was rewarding, since the animals quickly learned to press the lever and did not become satiated. Olds and colleagues (1956) found that the self-stimulation effect could be blocked by pharmacological substances such as chlorpromazine, which was later shown to block dopamine receptors. Only later, the relevance of the dopamine system to intracranial self-stimulation was demonstrated through lesions of the medial forebrain bundle (Olds & Olds, 1969) containing dopaminergic axons (Corbett & Wise, 1980). These findings first implicated the dopamine system in reward and motivation. In addition, it was recognized early on that the movement impairments in dopamine depletion-induced parkinsonism were problems of movement invigoration rather than an inability to move (Marshall et al., 1976), suggesting a role of dopamine in motivation. The self-stimulation findings also led to the "anhedonia hypothesis" of dopamine (Wise, 1985), which poses that dopamine is mediating the experience of pleasure, an account that was later largely revoked due to mounting evidence against it (Berridge, 2012; Wise, 2004).

The time course of dopaminergic activity in the context of reward and motivation was first investigated by Phillips and Olds (1969), who found that neurons in the ventral tegmental area of rats showed phasic increases in firing rate in response to salient stimuli, specifically tones paired with lever presses for food or water rewards, but not unpaired tones. This led the authors to the conclusion that the firing of the neurons depended on motivation. A subsequent series of studies by Schultz and colleagues investigated the firing patterns of midbrain dopamine neurons in nonhuman primates in different behavioral paradigms (Ljungberg et al., 1992; Mirenowicz & Schultz, 1994, 1996; Romo & Schultz, 1990; Schultz, 1986; Schultz et al., 1993; Schultz & Romo, 1990) and showed that dopamine neurons responded to rewards and reward-predicting cues. These experiments culminated in the dopamine reward prediction error hypothesis, which states that dopamine signals the difference between predicted and actual reward (Schultz et al., 1997). This hypothesis, which will be described in detail below, greatly influenced today's understanding of the role of dopamine in reward, motivation, learning, and decision-making.

2.3 Dopamine in learning and decision-making

Of the motor, cognitive and reward-related functions of dopamine described above, reward and motivation and their involvement in learning and decision-making are probably the most-researched aspects. This may be due to the close correspondence of dopamine function to the detailed mathematical descriptions of learning in reinforcement learning theory and the central relevance of learning and decision-making for cognition in general.

2.3.1 The dopamine reward prediction error hypothesis

The link between reinforcement learning theory and dopamine function, and thus a link between machine learning and biological learning, was established when Schultz and colleagues (1997) reviewed a large number of previous studies of electrophysiological recordings in dopamine neurons in nonhuman primates (Ljungberg et al., 1992; Mirenowicz & Schultz, 1994, 1996; Romo & Schultz, 1990; Schultz, 1986; Schultz et al., 1993; Schultz & Romo, 1990) and interpreted them in the context of reinforcement learning. During initial training, dopamine neurons showed phasic responses to rewards, while after Pavlovian learning, the neurons responded to reward-predicting cues and responded less to rewards. When these results were regarded in the light of temporal difference learning, it became clear that both the response to unpredicted rewards in the untrained state and the response to reward-predicting cues in the well-trained state could be interpreted as reward prediction errors (Schultz et al., 1997). Since the cue had become fully predictive of the reward, it was able to elicit a reward prediction error when presented at an unexpected timepoint. Just as predicted by the theory, the neurons stopped responding to the reward itself, once it became fully predicted by the cue. When a cue-predicted reward was omitted, the dopamine neuron response was below baseline, also in accordance with a negative prediction error posited by the model. There is now a large body of evidence in line with the dopamine reward prediction error hypothesis (Glimcher, 2011; Watabe-Uchida et al., 2017). The hypothesis of a role of dopamine in learning via reward prediction errors is also in line with dopamine-related plasticity mechanisms in the striatum (Reynolds et al., 2001; Yagishita et al., 2014). Yet, there are challenges to the hypothesis (Berke, 2018; Coddington & Dudman, 2019; Dayan & Niv, 2008), some of which will be discussed below.

Modern tools for the cell type-specific and projection-specific measurement and manipulation of neural activity in rodents with unprecedented spatiotemporal precision greatly aided the experimental investigation of the dopamine reward prediction error hypothesis. For example, using optogenetics for cell-type specific stimulation (as described in section 2.2.2), it was confirmed that activation of dopamine neurons was positively reinforcing, as shown by induced place preference (Tsai et al., 2009), and inactivation of dopamine neurons was negatively reinforcing, as shown by induced place aversion (Tan et al., 2012). The notion of dopamine as a teaching signal in form of a reward prediction error was experimentally tested using optogenetics in a so-called blocking paradigm (Steinberg et al., 2013). Blocking describes a phenomenon in Pavlovian conditioning, when an association between a conditioned stimulus A and an unconditioned stimulus will not be learned, if the conditioned stimulus A is presented together with another conditioned stimulus B that had already been associated with the unconditioned stimulus. Steinberg and colleagues (2013) showed that

activating dopamine neurons during the reward (i.e., unconditioned stimulus) led the animals to overcome the blocking effect and learn an association between a blocked conditioned stimulus and the unconditioned stimulus, presumably by adding a reward prediction error. Phasic dopamine activity therefore has been shown to induce learning via exogenous stimulation, providing causal evidence for the dopamine reward prediction error hypothesis. These findings are in line with early lesion studies that demonstrated the role of dopamine neurons in self-stimulation experiments (Corbett & Wise, 1980). Importantly, the unblocking effect of ventral tegmental area dopamine neuron stimulation (Steinberg et al., 2013) was replicated by Keiflin and colleagues (2019), although they found that stimulation of substantia nigra dopamine neurons did not produce the unblocking effect, but still supported reinforcement, which indicated heterogeneity among dopamine neurons. Further, Coddington and Dudman (2018, 2019) pointed out that mostly supraphysiological amounts of stimulation were used in experiments showing exogenous stimulation-induced behavior. They showed that stimulation in the natural physiological range is sufficient to produce learning, but not to directly invigorate action.

2.3.2 Debates around the reward prediction error hypothesis

Despite the large body of evidence in support of the reward prediction error hypothesis of dopamine function (Glimcher, 2011; Watabe-Uchida et al., 2017), there also remain a number of debates, all centering around the questions of how well dopamine signals can be reduced to reward prediction error and how other aspects, most importantly motivation and movement, can be incorporated into the hypothesis. There are also alternative accounts such as the incentive salience theory of dopamine function (Berridge, 2012; Berridge & Robinson, 1998) and other accounts that incorporate both value and salience (Bromberg-Martin et al., 2010; Kutlu et al., 2021).

One of the oldest debates around dopamine and reward learning is the question whether dopamine signals are used for learning or for motivation or vigor (Berke, 2018; Berridge & Robinson, 1998). In many studies of dopamine measurements and manipulations, dopamine signals can be interpreted equally well as teaching signals used for learning or as directly invigorating actions (Berridge, 2012). In addition, many neuropsychiatric symptoms related to dopamine functioning, including akinesia in Parkinson's disease, can be regarded as a motivational dysfunction. Even anhedonia symptoms in depression, that is, an apparent inability to derive pleasure from rewards, are usually more a problem of behavioral activation and drive than a deficit of hedonic experience itself (Treadway & Zald, 2011). The incentive salience theory (Berridge & Robinson, 1998) was put forward as opposing both the proposed role of dopamine in the experience of pleasure (Wise, 1985) as well as the role of dopamine in reward learning via prediction errors (Schultz et al., 1997), and instead holds that dopamine assigns incentive salience, a distinct component of motivation, to reward-related stimuli. Thus, the incentive salience theory deems dopamine to cause "wanting", as opposed to "liking" (experience of pleasure). While there is now largely a consensus that dopamine is not mediating pleasure, the question whether dopamine is responsible for learning or for motivation, or a complex mixture of the two, remains an open debate (Berke, 2018; Salamone & Correa, 2012). Recent observations of ramping activity of dopamine neurons during the anticipation or approach of reward

have sparked new interest in the discussion about the role of motivation in dopamine signals. Some authors regard dopamine ramps as motivational signals (Berke, 2018; Guru et al., 2020; Hamid et al., 2016; Howe et al., 2013; Mohebi et al., 2019), while others argue that ramps emerge as a property of reward prediction errors under certain conditions with specific shapes of value functions (Farrell et al., 2021; Gershman, 2014; Kim et al., 2020). Pan and colleagues (2021) recently found that stimulation of dopamine neurons in the ventral tegmental area of mice instead of natural rewards in Pavlovian learning produced conditioned approach behavior, but replacing reward-predicting cues with stimulation did not produce approach. Thus, dopamine stimulation was sufficient for reinforcement, but insufficient for reward prediction. The authors interpreted this as evidence against the incentive salience account, since dopamine as a cue replacement did not cause “wanting”, and emphasized the role of dopamine cue responses in learning rather than motivation.

A related question is whether dopamine neurons encode reward prediction error or the prediction itself, that is, value (or motivational value). Reward prediction error and value are hard to separate experimentally and many studies are not designed to do so. Reward prediction errors indicate a temporal difference and are approximately the derivative of values (Gershman, 2014; Kim et al., 2020). Thus, if dopamine signals are examined as changes relative to a baseline, which is commonly the case, value signals can look like reward prediction error signals. A series of recent optogenetic experiments demonstrated that cue-evoked dopamine signals in a Pavlovian learning paradigm represented reward prediction errors rather than values and dopamine stimulation did not add value to cues when associations were learned (Maes et al., 2020; Sharpe et al., 2020). Overall, this topic is still debated. While some authors suggest a differentiation across timescales and argue that reward prediction errors could be coded by fast phasic dopamine signals and motivational value could be coded by slower tonic (including ramping) dopamine signals (Berke, 2018; Niv et al., 2007; Schultz, 2007), others suggest that ramping and tonic signals as well as phasic signals, can be parsimoniously explained as reward prediction errors (Kim et al., 2020).

A significant challenge for both learning and motivation accounts of dopamine is the role of movement-related responses of dopamine neurons (Coddington & Dudman, 2019). Ever since the discovery of the role of dopamine in Parkinson’s disease, movement-related dopamine functions have been investigated. A number of recent studies has shown phasic dopamine signals to be related to various aspects of self-initiated movements, also in the absence of sensory cues (da Silva et al., 2018; Dodson et al., 2016; Hamid et al., 2016; Hughes et al., 2020; Lee et al., 2019; Syed et al., 2016). Movement-related signals are challenging for both reward prediction error framework and incentive salience theory, since neither can explain dopamine signals in response to self-initiated movements that are not externally motivated (Coddington & Dudman, 2019). Coddington and Dudman (2019) suggested a framework which incorporates action-related signals in reinforcement learning and poses that dopamine neurons encode action initiation in order to assign credit to successful actions and learn from them. Related to this, Coddington and colleagues (2021) demonstrated that behavioral changes during Pavlovian learning could be well explained by a policy learning model, a type of reinforcement learning algorithm that learns from errors in performance (as opposed to learning from reward prediction errors as in classical value learning). This suggests that action-related dopamine signals

could be incorporated in reinforcement learning through policy learning models in addition to (and not replacing) value learning models, a combination that has been successful in artificial learning (Silver et al., 2017).

Finally, several studies have found that a subset of dopamine neurons in response to aversive stimuli show increased firing, and not, as the reward prediction error theory would predict, decreases in firing (de Jong et al., 2019; Lammel et al., 2011; Matsumoto & Hikosaka, 2009). Increased firing to both appetitive and aversive stimuli can be interpreted as a salience signal, indicating behaviorally relevant events regardless of valence. This has led to alternative theories that propose distinct dopamine systems for value and salience (Bromberg-Martin et al., 2010) or suggest that dopamine (in the nucleus accumbens) exclusively encodes salience and only mimics reward prediction errors in some contexts (Kutlu et al., 2021).

2.3.3 Heterogeneity in dopamine neurons

As described above, several recent findings of dopamine signals correlated to different aspects of motivation (Mohebi et al., 2019), movement (Coddington & Dudman, 2019), aversion (de Jong et al., 2019), and salience (Menegas et al., 2017), as well as older theoretical debates (Berridge & Robinson, 1998), have challenged the view that dopamine uniformly conveys reward prediction errors to its projection targets (Lerner et al., 2020). While the main competing theories (e.g., learning versus motivation) try to explain the same reward-related aspects, recent findings of various aspects encoded by dopamine neurons that are in conflict with either theory (e.g., movement, aversion) revealed heterogeneity in populations of dopamine neurons. Engelhard and colleagues (2019) reported that dopamine neurons in the ventral tegmental area encoded several behavioral variables in a decision-making task, including reward history, trial accuracy, kinematics, and spatial position. They found that some neurons encoded more than one of these behavioral variables, and many neurons additionally encoded reward prediction errors. In contrast, movement-coding dopamine neurons in the substantia nigra seem to be more distinct from those encoding reward prediction error (da Silva et al., 2018; Howe & Dombeck, 2016). Further heterogeneity among dopamine neurons has been shown by Dabney and colleagues (2020), who found that the dopamine neurons in the ventral tegmental area that they analyzed all encoded reward prediction errors, but showed different degrees of optimistic or pessimistic predictions. This form of distributional coding had been shown to be beneficial in artificial learning and indeed seems to be implemented in biological systems as well. Another variable reported to modulate dopamine signals is decision confidence (Lak et al., 2017; Lak, Okun, et al., 2020), although some authors have argued that confidence correlations emerge from reward prediction errors (Tsutsui-Kimura et al., 2020).

Overall, dopamine neurons do not seem to encode a scalar reward prediction error and do not uniformly signal it to different targets, as maybe once assumed, but instead encode a variety of behavioral variables, some of which have been incorporated in the reward prediction error framework (Kim et al., 2020; Tsutsui-Kimura et al., 2020), while others still pose a challenge (Coddington & Dudman, 2019; Lerner et al., 2020).

2.3.4 Functional differences between dopaminergic projection targets

Some of the heterogeneous aspects that midbrain dopamine neurons encode may be grouped by the location of the cell bodies in the midbrain or by their projection targets. Dopamine neurons projecting to dorsal striatal subregions originate mainly in the substantia nigra, while dopamine neurons projecting to the ventral striatum originate mainly in the ventral tegmental area, but there are also connections that deviate from these paths (Björklund & Dunnett, 2007; Poulin et al., 2018). Thus, functional grouping of dopamine neurons may be easier according to projection targets than according to locations in the midbrain origin regions. There is evidence that different dopaminergic projections convey distinct signals and serve distinct functions.

The major dopaminergic projection target, the striatum, is an essential structure for reward learning. It receives inputs from thalamus and cortex and sends outputs via relay stations in two main pathways that oppositely modulate actions by either promoting or suppressing movement (Cox & Witten, 2019). Dopamine in the striatum is assumed to modify synaptic plasticity of glutamatergic inputs (e.g., from the cortex), when they are co-active with dopamine neurons within a narrow time window during unexpected rewards (Reynolds et al., 2001; Yagishita et al., 2014). At least three functional subregions have been identified in the striatum of rodents, non-human primates, and humans, mostly based on their involvement in different components of associative learning behavior (Redgrave et al., 2010). These include the dorsomedial striatum (DMS; roughly primate caudate), the dorsolateral striatum (DLS; roughly primate putamen), and the ventral striatum (VS; mainly comprised of the nucleus accumbens). The dorsomedial (or “associative”) striatum is thought to be important for goal-directed behavior based on action-outcome associations, while the dorsolateral (or “sensorimotor”) striatum is thought to important for habitual behavior based on stimulus-response associations (Balleine & O'Doherty, 2010; Cox & Witten, 2019; Redgrave et al., 2010; Yin & Knowlton, 2006; Yin et al., 2004; Yin et al., 2005). The VS is thought be involved in Pavlovian learning based on stimulus-outcome associations (Balleine & O'Doherty, 2010). Several findings have led to the view that the DMS mediates flexible, goal-directed behavior during early learning, with a shift towards engagement of the DLS during late learning and habit formation (Thorn et al., 2010; Yin et al., 2009). The striatal subregions are also implicated in so-called “actor/critic” reinforcement learning models that break down learning into two modules, an “actor” that selects actions according to a learned policy and a “critic” that learns value functions and evaluates on the basis of reward prediction errors (Barto et al., 1983; Sutton & Barto, 1998). In neurobiological implementations of the actor/critic model, the dorsal striatum acts as the actor, while the VS acts as the critic (Bornstein & Daw, 2011; Botvinick et al., 2009; Joel et al., 2002; O'Doherty et al., 2004). In theory, dopamine signals prediction errors to both areas, to the VS for value learning and to the dorsal striatum for policy learning, but so far there is little empirical evidence for this general hypothesis (Sutton & Barto, 2018). Other theories suggest that VS, DMS, and DLS are parallel and hierarchical reinforcement learning modules operating on different timescales to produce actions (Ito & Doya, 2011).

Striatal subregions receive different cortical inputs (Hunnicutt et al., 2016). Dopamine neurons with different striatal projection targets also receive distinct input patterns (Beier et al., 2015; Lerner et al., 2015) and show different burst firing properties (Farassat et al., 2019). Molecular grouping of

midbrain dopamine neurons has revealed remarkable specificity of distinct projection patterns among the molecularly defined subgroups (Poulin et al., 2018). It is therefore not surprising that there are also functional differences between dopamine signals in the striatal subregions.

Consistent with the proposed specializations of striatal subregions, dopamine in the DMS is required for cognitive flexibility in reversal learning (Grospe et al., 2018), dopamine in the DLS is required for habit formation (Faure et al., 2005), and dopamine in the VS is required for Pavlovian learning (Darvas et al., 2014). There is some evidence for a medial-to-lateral shift in dopamine release patterns during habit formation in rats using cocaine, with dopamine in the ventromedial striatum decreasing and dopamine in the DLS increasing from initial use to habitual use (Willuhn et al., 2012). Related to the specialization of striatal subregions in instrumental behavior, Hamid and colleagues (2021) recently reported wave-like activity patterns of dopamine axons across the striatum, propagating from DMS to DLS when rewards required instrumental actions, and from DLS to DMS, when rewards were independent of actions. They suggested that the dopamine waves may provide a mechanism for credit assignment to different striatal subregions depending on task demands and the extent of instrumental agency involved. Dopamine axonal activity in the VS was shown to faithfully represent reward prediction errors during Pavlovian learning (Menegas et al., 2017).

Further functional differences in dopamine activity across the VS, DMS and DLS have been identified in different behavioral paradigms (Howe & Dombeck, 2016; Parker et al., 2016; Tsutsui-Kimura et al., 2020; Wei et al., 2021). Tsutsui-Kimura and colleagues (2020) reported surprising similarity in dopamine release patterns during instrumental behavior across striatal subregions. One difference was a positively shifted reward prediction error pattern in the DLS compared to the other striatal subregions, lacking negative prediction errors. The authors interpreted this finding as support of the role of the DLS in habitual behavior, since theoretically a positively shifted reward prediction error reinforces action through repetition independently of the outcome. Wei and colleagues (2021) very recently showed that dopamine signals exhibited successively faster spontaneous fluctuations from the VS to the DMS to the DLS, with corresponding dynamics of integrating previous rewards and discounting future rewards. These different dynamics, according to the authors, could act as a mechanism for different time frames of specialized aspects of decision-making in the striatal subregions.

A fourth striatal subregion, the tail of the striatum (most posterior part of the dorsolateral striatum), has recently been shown to be distinct from the other subregions. Dopamine neurons projecting to the tail of the striatum have been shown to form a distinct anatomical (Menegas et al., 2015) and molecular (Poulin et al., 2018) subgroup and dopamine released in the tail of the striatum has been shown to signal salience (Menegas et al., 2017), reinforce avoidance (Menegas et al., 2018), and mediate hallucination-like perception in mice (Schmack et al., 2021).

Alongside the striatum, another major projection target of dopamine neurons is the prefrontal cortex. In primates, a medial-to-lateral topography can be identified in dopaminergic cortical projections. Neurons in the substantia nigra in the lateral midbrain primarily project to the lateral prefrontal cortex, while neurons in the ventral tegmental area in the medial midbrain primarily project to the medial prefrontal cortex (Ranganath & Jacob, 2016). This anatomical distinction is assumed to

be accompanied by a difference in information coding, with the medial projection being involved in motivational value signaling and the lateral projection being involved in cognitive salience signaling (Bromberg-Martin et al., 2010; Matsumoto & Hikosaka, 2009; Matsumoto & Takada, 2013). Since rodents do not have a granular lateral prefrontal cortex (Laubach et al., 2018), cortical dopaminergic projections in rodents mainly target medial prefrontal cortical areas. Therefore, the functional aspects of the mediolateral cortical projection gradient do not translate well from primates to rodents.

As mentioned in section 2.2.2, it has been suggested that dopamine release in striatal target regions can be modulated independently from somatic activity in the origin region (Mohebi et al., 2019), potentially through local modulation of release by cholinergic interneurons (Kosillo et al., 2016; Threlfell et al., 2012). Therefore, direct measurements of dopamine release in the target region, that is, in the region where dopamine has its effect as a neuromodulator, can reveal functional differences between projection targets that may not be detectable in dopaminergic cell bodies.

2.4 Dopamine signaling during task acquisition

Despite several open questions and debates around different functions of dopamine, its central role in learning is widely accepted, be it through learning of values and reward prediction errors (Schultz et al., 1997), salience and aversive signals (Bromberg-Martin et al., 2010; Kutlu et al., 2021), performance errors and policies (Coddington & Dudman, 2019; Coddington et al., 2021), or in conjunction with motivation (Berke, 2018). Yet, studies examining dopamine signals during the initial acquisition of learned associations, especially in behavioral tasks going beyond Pavlovian or simple instrumental associations, are surprisingly rare.

Due to the abundance of experimental techniques to measure and manipulate neuronal activity in animal models, many findings regarding dopamine function are based on behaviors that are commonly used in these animal models. In addition, the principles of learning theory originate from animal research of behaviorists in the twentieth century. While many aspects can be translated to humans, most phenomena and fundamental theories are still strongly influenced by the available behavioral assays to study learning in non-human animals. Due to the limited cognitive abilities of, for example, rodents, which are widely used, many aspects of the role of dopamine in learning are best researched in the domain of Pavlovian learning or simple instrumental learning. Instrumental learning paradigms are most commonly based on a so-called “bandit” task, where the subject learns to perform an instrumental action to obtain a reward, such as pressing a lever or choosing an arm in a T-maze. In a bandit task, the reward probabilities of different options (e.g., two levers, or two arms in a maze) are varied and learning is defined as corresponding adjustments in response rates for high-reward versus low-reward options. Based on the traditional classification of instrumental associations (Fig. 1), either action-outcome associations (goal-directed behavior), or stimulus-response associations (habitual behavior) are examined. Few studies in rodents have investigated more complex behaviors that involve different responses to different cues, that is, require the animal to acquire multiple associations between stimuli, actions, and outcomes, and to make decisions depending on the current state or context indicated by instruction cues.

Furthermore, most studies of instrumental learning examine animals in a well-trained state and define learning as updating of probabilistic reward values, rather than acquisition of an initial association. One reason for this may be the relatively late availability of chronic measurement tools (mostly fluorescent imaging techniques), which are required to track changes in neuronal activity during (often slow) learning processes.

2.4.1 Dopamine signals during Pavlovian learning

During Pavlovian learning, the phasic bursts of dopamine neurons shift from responding to the reward during early learning to responding to the reward-predicting cue after learning (Mirenowicz & Schultz, 1994). This transfer in dopamine transients is predicted by reinforcement learning models that model dopamine signals as reward prediction errors (Schultz et al., 1997). Temporal difference reinforcement learning models also predict a gradual temporal backward shift from the reward to the reward-predicting cue during learning (Schultz et al., 1997; Sutton & Barto, 1998). While the transfer itself has been observed numerous times both in the activity of dopamine neurons and in the activity of

their axons in the striatum (Flagel et al., 2011; Menegas et al., 2017; Mirenowicz & Schultz, 1994; Pan et al., 2005), the gradual temporal shift has been shown less clearly until recently (Amo et al., 2020). Amo and colleagues (2020) found that a gradual shift could be observed in dopamine axons in the ventral striatum under certain favorable conditions, including the use of calcium imaging with relatively slow kinetics (compared to electrophysiology) and the recording of average population activity instead of single cells. Further, temporal difference learning models showed that the occurrence of a gradual shift depended on the eligibility trace parameter λ, which determines the number of time steps eligible for modification by reward prediction errors, that is, how far the model looks back in time during credit assignment (Amo et al., 2020; Pan et al., 2005).

The temporal difference model assumes a single system for reward prediction error responses to both the reward and the reward-predicting cue, and a simultaneous decrease of the reward response and increase of the cue response during learning (Sutton & Barto, 1998). Menegas and colleagues (2017) found different timescales for the changes in reward responses and cue responses of dopamine axons in the ventral striatum during Pavlovian learning. The reward response decreased together with an increase in behavioral signs of learning (anticipatory licking), while the cue response increased much slower over the course of several sessions. A similar lag of the dopamine cue response relative to learning behavior has been observed in other studies of Pavlovian learning (Coddington & Dudman, 2018; Coddington et al., 2021). Coddington and Dudman (2018) found that in their Pavlovian learning paradigm, reward prediction error correlates in ventral tegmental area and substantia nigra dopamine neurons only emerged after learning, which according to their interpretation can be attributed to the fact that reward prediction error correlates are a consequence of temporal integration of reward expectations related to reward-predictive cues and appetitive action initiation. Their data also suggested that dopamine cue responses and reward response could emerge independently. These dissociations question the causal role of reward prediction error correlates in initial learning of associations. Pan and colleagues (2021) recently probed the causal roles of predictive versus reinforcing aspects of dopamine reward prediction errors by replacing either reward-predictive cues or rewards with optogenetic stimulation of dopamine neurons during Pavlovian learning. They found that dopamine stimulation replacing rewards (i.e., reinforcement) elicited behavioral signs of learning, while dopamine stimulation replacing reward-predictive cues (i.e., prediction) did not. These results indicate a causal role of dopaminergic reinforcement in learning and question the role of dopaminergic reward prediction in learning.

In contrast to the lags between reward prediction error signatures and behavior during the acquisition of novel Pavlovian associations, Menegas and colleagues (2017) found that during repeated learning using novel stimuli, reward prediction error signatures developed much faster than during initial acquisition, although reward response and cue response still changed in the same order, that is, the cue response evolved slower compared to changes in the reward response. Many studies have investigated learning in Pavlovian paradigms by studying animals in the well-trained state. In these paradigms, learning is usually not defined as the acquisition of novel associations in naïve animals, but as the updating of reward values, which is experimentally induced by changes in reward contingencies and manipulating reward magnitudes or probabilities. Dopamine reward prediction

errors in Pavlovian paradigms scale with the expected value of reward, as determined, for example, by reward magnitude or reward probability (Fiorillo et al., 2003; Tobler et al., 2005). Reward prediction error signatures in dopamine neurons quickly reflect expected values upon the introduction of novel stimuli (Lak et al., 2016). Interestingly, Lak and colleagues (2016) found that even though adaptations of dopamine reward prediction error signals were fast, behavioral signs of learning reflecting reward probabilities emerged earlier than corresponding signatures in the dopamine signals.

By and large, theoretical predictions of reinforcement learning models stating that dopamine as a reward prediction error transfers from the reward to the cue, and does so in a gradual way, have been confirmed in experimental studies using Pavlovian learning paradigms. Lags between behavioral signs of learning and reward prediction error signatures that cannot be easily explained by temporal difference learning models have been observed both during initial learning and during repeated learning or updating of probabilistic reward values.

2.4.2 Dopamine signals during instrumental learning and decision-making

Compared to studies of Pavlovian learning, fewer studies have looked at dopamine signals during instrumental learning and decision-making, especially during the initial acquisition of associations in naïve animals. When rewards are not fully (or probabilistically) predicted by reward-predicting cues as in Pavlovian learning, but an instrumental action or a specific choice between multiple actions is required to obtain a reward, the reward prediction error signature inherently gets more complicated. The predicted value of a behaviorally relevant stimulus now not only depends on an externally set reward probability or magnitude, but also on the subject's knowledge of the task and associations between stimuli, required actions, and outcomes. Thus, a simple transfer of reward prediction error signals from the reward to the reward-predicting cue as in Pavlovian learning is not to be expected.

While dopamine neurons encode the expected value of reward-predicting cues during Pavlovian learning (Fiorillo et al., 2003; Tobler et al., 2005), when animals are given a choice between two options in instrumental tasks, dopamine neurons encode the expected value of the chosen option (Lak et al., 2016; Morris et al., 2006). Lak and colleagues (2016) studied prediction error signatures in monkey dopamine neurons both during Pavlovian learning (described above) and in a choice task with changing stimuli. Just like during Pavlovian learning, in the choice task dopamine neurons quickly adapted their response during the stimulus epoch to the reward probabilities of novel stimuli, again lagging behind behavioral preferences for cues according to expected reward values.

The dopamine response to the reward has been shown to scale with task performance in monkeys trained in a similar choice task (Hollerman & Schultz, 1998). In this early study by Hollerman and Schultz (1998), monkeys had to choose between rewarded and non-rewarded stimuli, which were exchanged with novel stimuli in blocks of trials. The animals quickly adapted to the novel stimuli over a few trials and performance increased. Together with an increase of correct choices, the dopamine reward response decreased, likely because rewards became better predicted and the reward prediction error decreased. How the dopamine response to stimuli emerged during increases in performance was not described in the study by Hollerman and Schultz (1998).

Many studies have looked at dopamine signals in animals performing decision-making tasks in the fully trained state, when associations are already acquired. One way to study learning during this consolidated behavior is to experimentally induce variability in reward magnitude or probability (Hamid et al., 2016; Lak, Okun, et al., 2020; Mohebi et al., 2019; Tsutsui-Kimura et al., 2020). Adapting behavior to changes in reward contingencies can be described as learning, although associations are already established.

Studies of this kind have found that dopamine reward prediction error signals correlate with subjective preference (Lak et al., 2014) and in addition to reward expectations incorporate the animals' belief states about decision confidence (Lak et al., 2017). Dopamine reward prediction error signals based on belief states have also been observed in Pavlovian learning tasks (Babayan et al., 2018; Starkweather et al., 2017). Lak and colleagues (2020) varied reward magnitude and perceptual difficulty in a visual decision-making task in mice and found that dopamine neurons in the ventral tegmental area encoded confidence-dependent predicted value and prediction error during the cue and reward epochs, respectively. When they optogenetically manipulated the activity of dopamine neurons, only manipulating reward prediction errors during the reward epoch impacted behavior. As in the study of Pavlovian learning described above (Pan et al., 2021), this suggests a causal role of the reinforcing properties of dopamine reward prediction errors in learning, but no causal role of the predictive properties. Taken together, the findings mentioned above indicate that reward prediction error signatures during complex decision-making tasks depend on current state and behavioral strategy of the animal and are not purely determined by experimentally set reward contingencies.

Overall, while dopamine signals have been well-described in Pavlovian learning (Menegas et al., 2017), learning as adaptations to novel stimuli in trained animals performing decision-making tasks (Hollerman & Schultz, 1998; Lak et al., 2016), and learning as adaptations to changes in reward probabilities or magnitudes in decision-making tasks (Lak, Okun, et al., 2020; Tsutsui-Kimura et al., 2020), it is not well understood how dopamine signals evolve during learning defined as acquisition of novel instrumental associations in naïve animals. It is especially elusive how the dopamine response to instruction cues develops during acquisition of associations when external reward probabilities are constant and variability in reward deliveries only arises from the behavioral performance and thus the learning state and association strength. Further, it remains an open question how the predictive aspect (during instruction cues) and the reinforcing aspect (during the outcome) of dopamine reward prediction errors evolve and interplay during learning of novel instrumental associations.

2.5 Present study

In this study, mice were trained in an auditory decision-making task with several rule switches, which required the mice to acquire associations between different instruction cues and actions to make correct choices in order to obtain rewards. To investigate the time course of dopamine signals in the main learning-related striatal subregions (the ventral, dorsomedial, and dorsolateral striatum) during stimulus, action, and outcome, and how these signals changed during the acquisition of associations, chronic fiber photometry using the fluorescent dopamine sensor dLight (Patriarchi et al., 2018) was performed while mice learned the task. The rationale behind this study can be summarized by three main aspects.

First, few studies have addressed the role of dopamine in the acquisition of decision-making tasks, while many studies examined dopamine in Pavlovian tasks (Coddington & Dudman, 2018; Menegas et al., 2017; Pan et al., 2021) or simple bandit tasks (Mohebi et al., 2019). In these behaviors, the receipt of a reward either requires no action at all or a simple instrumental action that does not depend on trial-by-trial rule-based decisions in response to instruction cues, but on slowly changing reward probabilities of available options. It is important to understand how dopamine signals are involved in the acquisition of rule-based associations in a decision-making task, since dopamine signatures may be different from the well-described signatures in Pavlovian and simple instrumental tasks when the delivery of a reward depends on a correct decision and not only on external reward probabilities. Therefore, in this study external reward probabilities were kept constant and the only element that varied across the learning process was the acquired association between instruction cue, required action, and desired outcome. The central question related to this aspect was how dopamine cue responses and reward responses evolve during the acquisition of instrumental associations. The dopamine reward response was expected to inversely scale with reward expectation, and thus task performance, as previously shown in a choice task (Hollerman & Schultz, 1998) and as predicted by the reward prediction error framework (Schultz et al., 1997). Analogously, the dopamine cue response could be scaled with reward expectation. In Pavlovian tasks, the dopamine cue response simply scales with reward expectation (Fiorillo et al., 2003; Tobler et al., 2005) and in instrumental tasks with varying external reward probabilities the dopamine cue response scales with reward expectations when animals are well-trained (Lak et al., 2016; Morris et al., 2006). Therefore, one hypothesis was that in the present task, the dopamine cue response should scale with task performance, since external reward probabilities were kept constant and only the state of the learned association (as indicated by task performance) determined the average reward probability, and thus reward expectation. The present experiments tested whether this hypothesis could be supported despite the aforementioned differences of the present behavioral task to previous studies.

Second, most studies of dopamine signals during decision-making did not examine the initial acquisition of associations, but studied animals in the fully trained state and defined learning as updating of probabilistic reward values (Lak, Okun, et al., 2020). Studying the initial acquisition of associations is important in order to better understand the role of reward prediction errors during learning of associations and how predictive aspects and reinforcing aspects of dopamine reward prediction errors contribute to the learning process (Coddington & Dudman, 2018; Pan et al., 2021).

Therefore, in this study dopamine signals were examined from the very beginning, when naïve mice started to learn the task, up until they reached plateau performance, and throughout further learning processes after rule switches. The central question related to this aspect was how the predictive aspect of dopamine reward prediction errors during reward-predicting cues and the reinforcing aspect of dopamine reward prediction errors during rewards contributed to learning. While temporal difference models assume a single system for reward prediction error responses to both the reward and the cue, which should evolve simultaneously during learning (Sutton & Barto, 1998), previous studies have reported lags between behavioral signs of learning and specific aspects of dopamine reward prediction error signatures in Pavlovian learning (Coddington & Dudman, 2018; Lak et al., 2016; Menegas et al., 2017) and repeated instrumental learning (Lak et al., 2016). Recent optogenetic studies found a causal role of dopaminergic reinforcement, but not dopaminergic prediction, in Pavlovian learning (Pan et al., 2021) and during learning as updating of probabilistic reward values (Lak, Okun, et al., 2020). The present experiments allowed to test whether the predictive and reinforcing aspects of reward prediction error signatures were equally or differentially correlated to behavioral signatures of the acquisition of associations during instrumental learning.

Finally, in order to elucidate potential differences between dopaminergic projection targets, direct fluorescent measurements of dopamine using fiber photometry were performed in several subregions of the striatum while mice learned the task. Imaging dopamine via a fluorescent sensor in the target region is advantageous for examining the effects of dopamine release per se, disentangling it from potentially different activity at the soma (Mohebi et al., 2019) and potential co-released glutamate (Mongia et al., 2019), which can have redundant effects on reinforcing behavior in the ventral striatum (Wang et al., 2017; Zell et al., 2020). The present experiments allowed to compare how dopamine cue and reward responses and the predictive and reinforcing aspects of reward prediction error signatures evolved in the ventral, dorsomedial, and dorsolateral striatum across learning. Based on proposed functional specializations of striatal subregions (reviewed in section 2.3.4) and previous comparisons of dopamine measurements across these subregions (Howe & Dombeck, 2016; Menegas et al., 2017; Parker et al., 2016; Tsutsui-Kimura et al., 2020; Wei et al., 2021), both differences and commonalities were expected.

3. Results

3.1 Multi-timescale behavioral analysis of decision-making task acquisition

To investigate the time course of dopamine signals across subregions of the striatum during decision-making and how these signals evolve during learning, I developed an auditory decision-making task with rule switches for head-fixed mice. The task was trained over several weeks. Chronic real-time measurements of dopamine were performed in three subregions of the striatum throughout the learning process. As a basis for the analysis of the dopamine data, a thorough behavioral analysis was conducted in order to identify behavioral strategies of mice performing the task.

3.1.1 Mice learn auditory decision-making task

Head-fixed mice were trained to associate an auditory instruction cue with a directed licking response in order to obtain a desired outcome. In each trial, the animals had to choose between two drinking spouts by licking one of the two spouts, which were presented after an auditory instruction cue was played from one of two speakers (Fig. 3A). The auditory cue indicated which of the two spouts would dispense a water reward upon a correct instrumental lick. The auditory cues were low-frequency or high-frequency band-pass filtered white noise sounds played from one of the speakers to the left or to the right of the animals. The two cue dimensions, frequency and location of the sound, were fully crossed, yielding four different conditions, which were presented in pseudo-randomized order for several hundred trial trials per session. An implicit task rule could be set according to one of the two instruction cue dimensions (Fig. 3B). Over the course of several sessions, the animals first learned to respond to the location of the sound, that is, a response was counted as correct, when the animals licked the left spout in response to a sound from the left speaker or the right spout in response to a sound from the right speaker, while the frequency of the sound was irrelevant. The relevant cue dimension was then switched from the location to the frequency of the sound (extra-dimensional rule switch) and finally switched within the dimension of frequency by completely reversing the rule (intra-dimensional rule switch).

Figure 3C shows a schematic of the trial structure. The auditory cue lasted one second, during which the animals could already prepare for their response and make a decision, but they could only perform the instrumental lick when the two licking spouts were moved into the reach of the tongue after the presentation of the cue. Upon a correct lick, a water reward was dispensed from the target spout and the non-target spout was retracted immediately. The animals consumed the reward within two seconds before the target spout was retracted. Upon a false lick, both spouts were retracted immediately and a time-out was triggered in addition to the regular inter-trial interval.

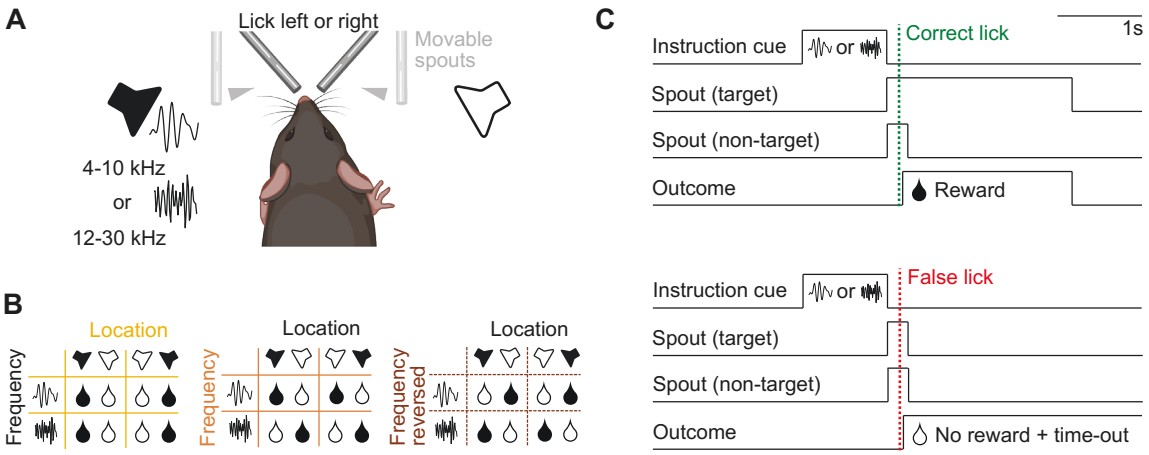


Fig. 3) Auditory decision-making task
A) Schematic of the animal head-fixed in the setup. Water rewards were dispensed from movable drinking spouts upon correct choices. Instruction cues were low frequency (4 to 10 kilohertz) or high frequency (12 to 30 kilohertz) noise sounds, played from one of two speakers to the left and to the right of the animal. Schematic of mouse and spouts created with Biorender.com and used with permission. **B)** Fourfold table schematic of the conditions and rule switches. Filled speaker symbols indicate the location of the sound (left or right) and filled water drop symbols indicate the side of the water reward in case of a correct response. **C)** Schematic of the trial epochs for correct trials (upper panel) and error trials (lower panel).

3.1.2 Mice learn different task rules

After habituation and pre-training to get accustomed to the licking setup (see *Methods*), mice started learning an implicit task rule by trial and error. Figure 4A depicts the average fraction of correct trials per session for one example mouse across the whole training period. The animals sequentially encountered three different task rules. The rules were switched after the animals had reached the criterion of at least 80 percent of correct trials overall and at least 60 percent of correct trials in each single condition (see *Methods*). Figure 4B shows the average performance trajectories for all animals (n = 26) across the three different task rules. Mice gradually learned the different task rules across several sessions. During the initial task acquisition, most animals started with an idiosyncratic preference for one of the two spouts and mainly licked at one of the two spouts during the first sessions (Fig 4C). This response bias gradually decreased across multiple sessions while the average performance increased. Mice reached criterion performance in the location task on average after 14.67 sessions (standard deviation 5.77 sessions). After acquisition of the first task rule, the rule was switched from the location of the instruction cue to the frequency of the instruction cue. Animals reached criterion performance in the frequency task on average after 5.67 sessions (standard deviation 2.44 sessions). Finally, the task rule was switched within the dimension of frequency by completely reversing the contingencies. Animals reached criterion in the frequency reversed task on average after 7.79 sessions (standard deviation 2.74 sessions). In the frequency reversed task, animals were trained until they reached asymptotic performance, as visually observed in the learning curves (Fig. 4B). Overall, the animals required significantly more sessions to reach criterion performance during the initial task rule – the location rule – compared to both frequency rules. Further, the animals required significantly more sessions to criterion during the frequency reversed rule compared to the frequency rule (Fig. 4E; see also *Appendix* Table A1 in section 6 for all statistical test results), most likely because they started the new task rule at a lower performance level during the frequency reversed task compared to the frequency task (Fig. 4B/D). The different reactions to the two rule switches are described in more detail below.

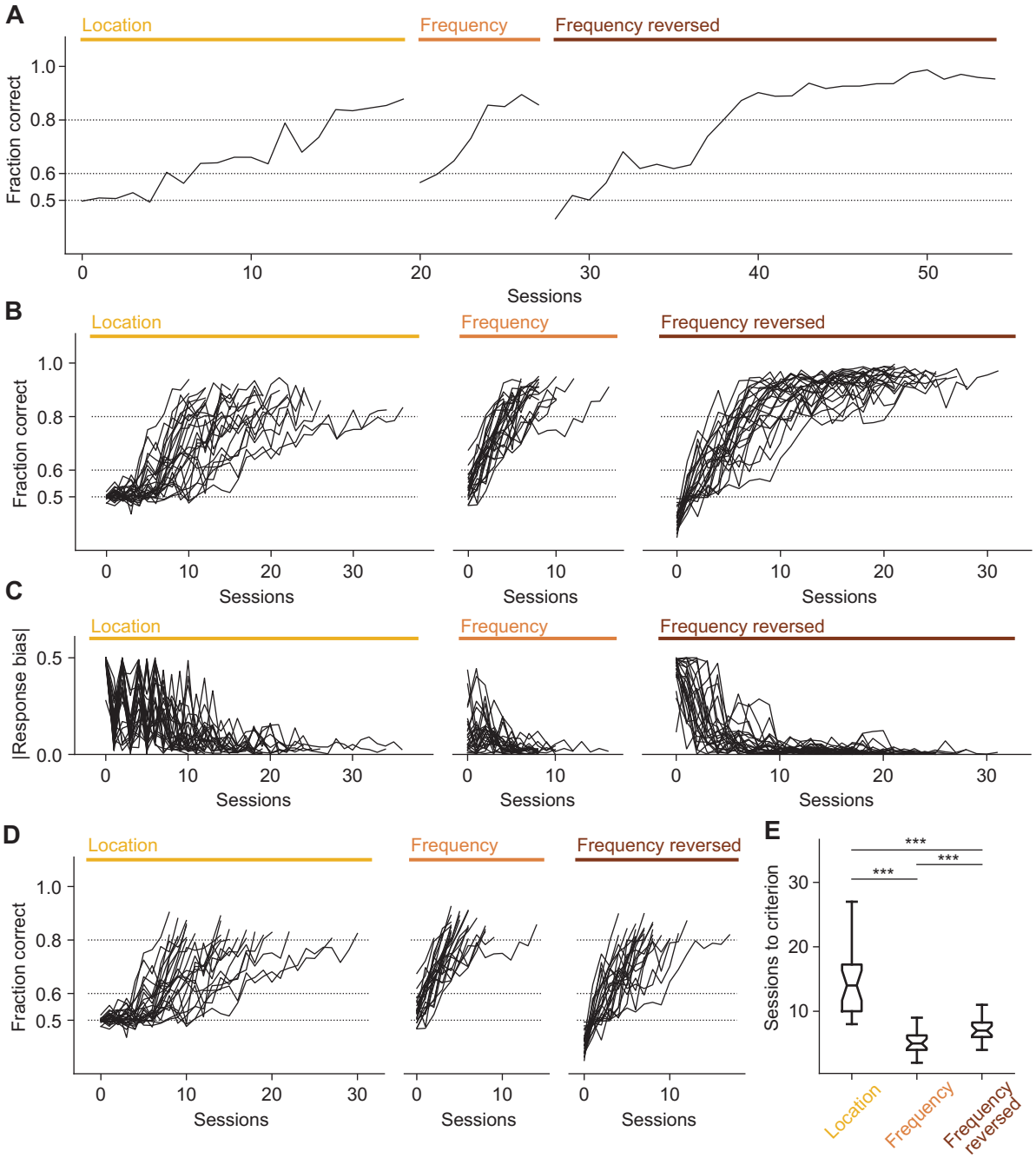


Fig. 4) Behavior during task acquisition
A) Average fraction of correct trials per session for one example animal across three task rules. **B)** Average fraction of correct trials per session and per animal for n = 26 animals across three task rules, aligned to the rule switches. **C)** Average absolute response bias (see *Methods*) per session and per animal for n = 26 animals. **D)** Same as B), but only up to the first session with criterion performance. **E)** Average number of sessions to criterion per task rule. Location vs. frequency: $p < 10^{-6}$; location vs. frequency reversed: $p < 0.001$; frequency vs. frequency reversed: $p < 10^{-6}$. *** indicates $p<0.001$. Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons. See also *Appendix* Table A1 in section 6 for all statistical test results.

Since the animals on average acquired the frequency rule faster than the location rule, the question arises whether mice had a general preference for the frequency rule over the location rule. In a pilot experiment, the order of the cue dimensions was varied and half of the animals started with the frequency rule. The number of sessions to criterion did not differ significantly across the two groups of animals with different starting rules (Fig. 5). It has to be noted that the two groups in the pilot experiment were trained under the same protocol and only the starting rule was varied, but the pilot protocol slightly differed from the final protocol in the main experiments (see *Methods* for experimental details). This pilot experiment gave no indication that the animals had an advantage for either of the two cue dimensions per se.

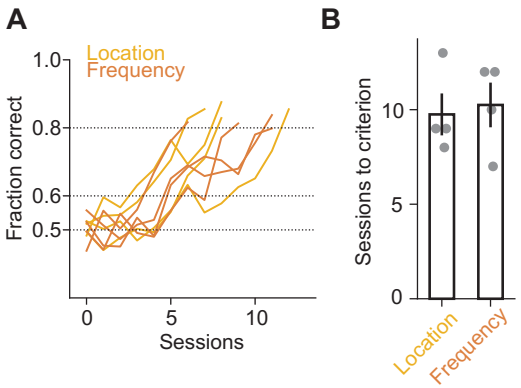


Fig. 5) No general preference for location or frequency
A) Average fraction of correct trials per session and per animal for two groups of animals (location rule vs. frequency rule). **B)** Average number of sessions to criterion per task rule. Bars show mean, error bars show standard error of the mean across animals. Permutation test, n = 4 animals (location) vs. n = 4 animals (frequency), p = 0.886.

In summary, mice reliably acquired the auditory decision-making task. Mice required the largest number of sessions to acquire the initial task rule and they required more sessions to acquire the frequency reversed task rule compared to the frequency task rule. This indicated that the animals reacted differently to the rule switch from location to frequency compared to the rule switch from frequency to frequency reversed. Note that the experimental setting was different for the two rule switches (see also section 3.1.3). After the rule switch from location to frequency the contingencies were switched only for two out of four conditions, while they were switched for all four conditions after the rule switch from frequency to frequency reversed (Fig. 3B).

3.1.3 Mice perseverate after rule switches

In order to further understand the differences in behavior during the different task rules, a closer look at intra-session changes during rule switch sessions is helpful. During the session after the rule was switched from the location to the frequency of the cue, in two out of four conditions the contingencies stayed the same (“stay” conditions), while they changed in the other two conditions (“switch” conditions). All animals first kept responding according to the location rule, that is, their responses were mostly correct in stay conditions while they were mostly incorrect in switch conditions (Fig. 6A), at least in the first 100 trials of the session (Fig. 6C). In the session following the rule switch from frequency to frequency reversed, mice again perseverated, that is, kept responding according to the previous rule during the first trials after the switch (Fig. 6B/C). Since in this switch session the contingencies changed for all conditions, all trials were switch trials and mice started out with virtually no correct trials and exhibited an average performance of below 0.5 fraction correct, at least during the first 100 trials (Fig. 6C). After around 100 trials of perseveration, most animals showed an average performance of around 0.5 fraction correct (Fig. 6B, upper panel). This performance was accompanied by a strong absolute response bias approaching the maximum value of 0.5 (Fig. 6B, lower panel), indicating that the animals licked only one of the two spouts, which they idiosyncratically preferred.

In summary, due to the different behaviors during rule switch sessions, the animals had different starting conditions for the two task rules after rule switches. Most animals started the frequency rule with a session average overall performance of above 0.5 fraction correct and did not show strong response biases (Fig. 4B/C). In contrast, most animals started the frequency reversed rule with an average performance of below 0.5 fraction correct and strong response biases, which many animals kept up for several further sessions before the performance increased and the response bias decreased again (Fig. 4B/C).

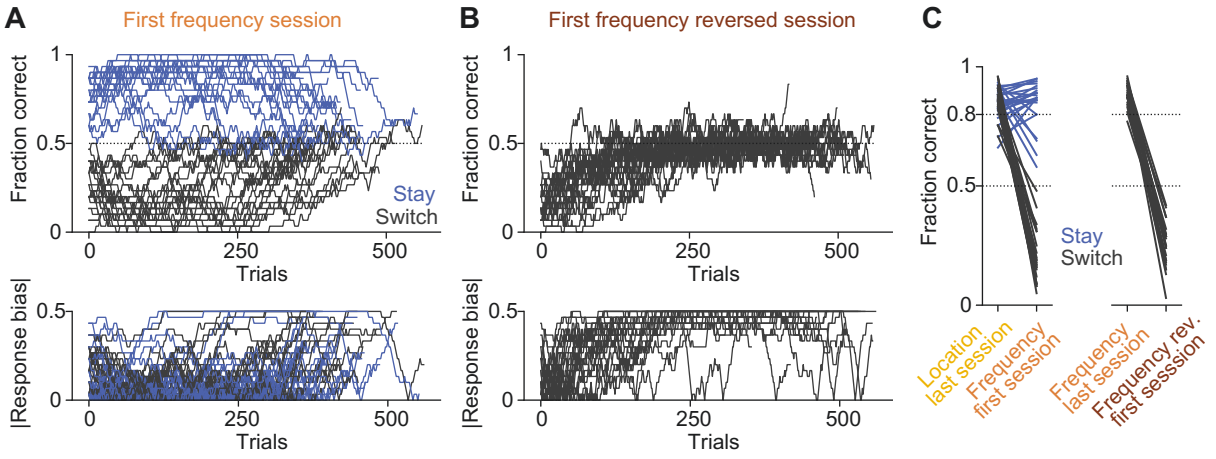


Fig. 6) Behavior after rule switches
A) 30-trial running average of fraction of correct trials (upper panel) and absolute response bias (lower panel) per animal in the first session of the frequency task, split by conditions in which the correct response stayed the same (“stay”) vs. conditions in which the correct response switched (“switch”). **B)** Same as A) for the first session of the frequency reversed task; all conditions were switch conditions. **C)** Summary of the average fraction of correct trials in the first 100 trials of the last session of the previous task rule and the rule switch session for the two different rule switches.

In every other behavioral session throughout the whole learning process, direct fluorescent imaging of dopamine was performed using fiber photometry. This experimental regime produced quasi-continuous data points of dopamine measurements across learning (Fig. 7A). In order to examine changes in dopamine across learning, the imaging sessions were split into three performance groups for further analysis (Fig. 7B). “Beginner” sessions included sessions, during which the animals showed an average session performance with a fraction of correct trials below 0.6. “Intermediate” sessions included sessions with an average performance of 0.6 fraction correct or above, but not criterion sessions, and “expert” sessions included all criterion sessions, that is, sessions with an average performance of 0.8 fraction correct or above overall and 0.6 fraction correct or above in each single condition (see also *Methods*). A summary of the performance levels across task rules again shows that the animals reacted differently to the two rule switches, with an average performance of above 0.5 fraction correct in beginner sessions during the frequency rule in contrast to an average performance of below 0.5 fraction correct in beginner sessions during the frequency reversed rule (Fig. 7B). Further, a summary of the average absolute response bias split by performance levels and task rules again shows that the animals started the location task and the frequency reversed task, but not the frequency task, with strong response biases (Fig. 7C).

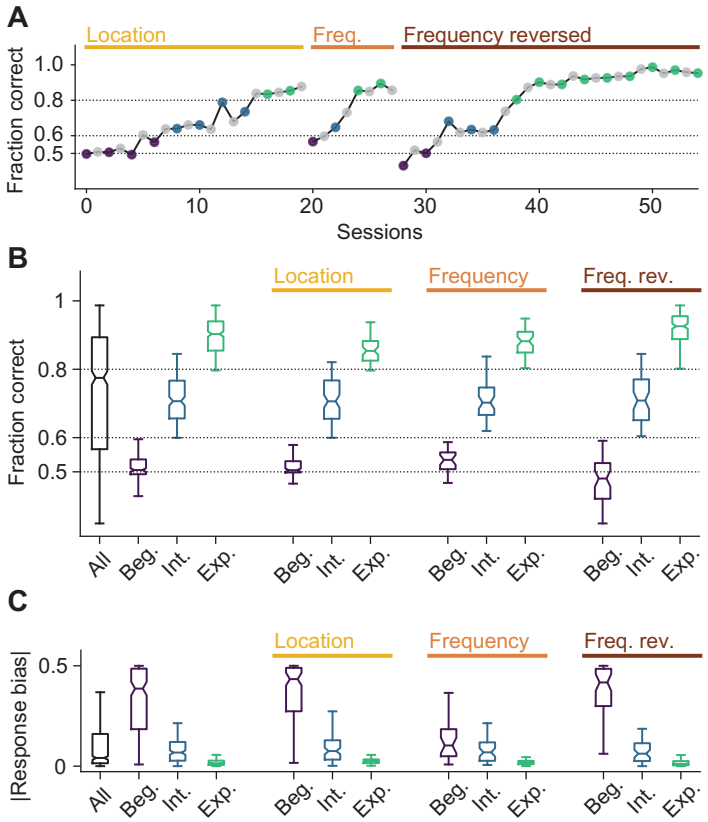


Fig. 7) Performance levels
A) Average fraction of correct trials per session for one example animal across three task rules. Gray dots indicate non-imaging sessions; other dots indicate imaging sessions (every other session), color-coded by performance levels “beginner”, “intermediate”, and “expert” (see B and C). **B)** Average fraction of correct trials per session for all imaging sessions, split by performance levels and task rules (n = 692 sessions; n = 193 beginner sessions, n = 190 intermediate sessions, n = 309 expert sessions). **C)** Average absolute response bias split as in B).

3.1.4 Session-based choice model reveals relevant choice predictors

The session-average trajectories of performance and response bias across performance levels and task rules indicated that mice used different strategies in their effort to perform the task. For example, mice exhibited strong idiosyncratic response biases in beginner sessions in the location task and in the frequency reversed task. Further, the intra-session trajectories of performance and response bias in sessions following rule switches indicated that mice used different strategies within a single session. For example, mice switched from perseveration on the previous rule to a response bias strategy in the frequency reversed rule switch session. In order to better characterize the different strategies and to further understand the influences on mouse choices and how they change during learning, logistic regression models were used to predict mouse choices. These choice models help to elucidate relevant behavioral variables that guide the animals' decisions (Busse et al., 2011). Logistic regression models were used to quantify the contribution of task variables and trial history-related variables to the animals' choice in each trial. Comparing different models by iteratively adding and removing regressor variables can reveal the relative contributions of these variables to the animals' choices (Akrami et al., 2018). To complement this analysis, changes in model coefficients, that is, weight trajectories across sessions and trials can be used to identify influences of different task-related variables on choice and how they change across learning (Roy et al., 2021).

To identify the relevant predictors and their relative influences across learning, a logistic regression model was first fit for each session. Models with different parameters were tested against each other using cross-validation, that is, by fitting coefficients on a subset of trials and testing predictions on an unused subset of trials (see *Methods*). The final session-based model (Fig. 8A) included the following predictors: an intercept, which represents a global response bias towards one of the two licking spouts; two predictors for the two instruction cue dimensions, which represent the strategy of using the location or the frequency of the auditory cue to guide the choice; and a trial history-related predictor representing the influence of the recent history of choices and rewards on the current choice. The trial history predictor in the final session-based model was "value difference", which was defined as the difference between the current reward rate for the left spout and the current reward rate for the right spout based on an exponentially weighted average over the last 10 trials (see *Methods* for details). It is important to note that the reward size was constant and the occurrence of a reward was not probabilistic but only contingent on the cue and the choice. Therefore, "value difference" indicated the value difference of the two spouts solely based on the recent history of choices and rewards. For example, if an animal only responded to the right spout for a certain number of trials, it would also receive rewards only on the right spout, resulting in a value difference between the spouts. A positive weight for the value difference predictor thus biased the choice of the model towards the currently preferred spout, based on the recent trial history. When the animal did not have a preference for either of the spouts, the value difference only indicated minor fluctuations in recent trial history due to the randomization of trials. The final session-based model only included predictors that if removed led to significant degradation of cross-validated prediction accuracy of the model (Fig. 8B, *Reduced model*s). The final model was compared to models with alternative predictors for trial history and to other models with additional predictors. Replacing or adding trial history or cue-related

predictors did not further improve cross-validated prediction accuracy or even degraded it (Fig. 8B). The alternative trial history predictors included the reward rates of the left and right spout (as two separate predictors instead of a combined value difference predictor), the history of choices in the previous three trials, and two predictors representing a “win-stay” and “lose-switch” strategy, respectively (see *Methods* for details). The additional variables included the previous trial choice and previous trial outcome and the interaction thereof, the previous choice and outcome and interaction thereof up to three trials back, the win-stay and lose-switch predictors, the interaction of the instruction cue predictors, and the history of the instruction cue up to three trials back. The final model was the most parsimonious model that showed the largest cross-validated prediction accuracy (Fig. 8B; individual results of this and all other statistical tests can be found in *Appendix* Table A1 in section 6). Two models with alternative trial history predictors exhibited similar prediction accuracy, but included two to three predictors for trial history, which was represented by only one predictor (value difference) in the final model. While the alternative trial history predictors explained a similar amount of variance in the choice data compared to the value difference predictor in the final model, they did not improve model performance when added to the final model, indicating that they did not have unique contributions (Fig. 8B). Thus, the trial history-related aspects could be accounted for equally well by different representations of past rewards and choices (i.e., value difference, reward rate on the left and right spout, or choice history of the previous three trials), but not by a win-stay or lose-switch strategy. The most parsimonious predictor (i.e., value difference) was chosen for the final model.

Averaged across all sessions, the cross-validated prediction accuracy indicated that a full model including cue-related and trial history-related predictors performed best compared to reduced models (Fig. 8C). In beginner-level sessions, when animals showed strong idiosyncratic side preferences, the (response) bias-only (i.e., intercept-only) model already explained a large amount of variance in the choice data compared to models with more predictors; the history-only performed significantly better than the bias-only model; the cue-only model did not perform better than the history-only model; and the full model performed significantly better than the cue-only model (Fig. 8C). These results indicated that in beginner sessions the animals exhibited strong response biases and used the recent history of choices and rewards, but not the instruction cue, to guide their choice. In intermediate sessions, the history-only model performed significantly better than the bias-only model; the cue-only model performed significantly better than the history-only model; and the full model performed significantly better than the cue-only model (Fig. 8C). These results suggested that in intermediate sessions the animals used the instruction cue to guide their choice while still relying on the recent history of choices and rewards. In expert sessions, the history-only model did not perform significantly better than the bias-only model, while the cue-only model performed significantly better than the history-only model; the full model did not perform significantly better than the cue-only model (Fig. 8C). This suggested that on average the recent history of choices and rewards did not significantly influence choice when the animals had fully learned the task and performed at expert level. Instead, when the animals were experts, they followed only the instruction cue. Note that prediction accuracy of the full model was lowest in intermediate sessions, indicating that during these sessions the fraction of choices not explained by the model was larger than during beginner or expert

sessions. The animals' strategies were captured well by response biases in beginner sessions and by following the instruction cue in expert sessions. In contrast, in intermediate sessions, when animal performance was most volatile and learning took place, the animals' strategies contained contributions of response bias, trial history-related aspects, and cue-related aspects, as well as additional aspects not captured by the model.

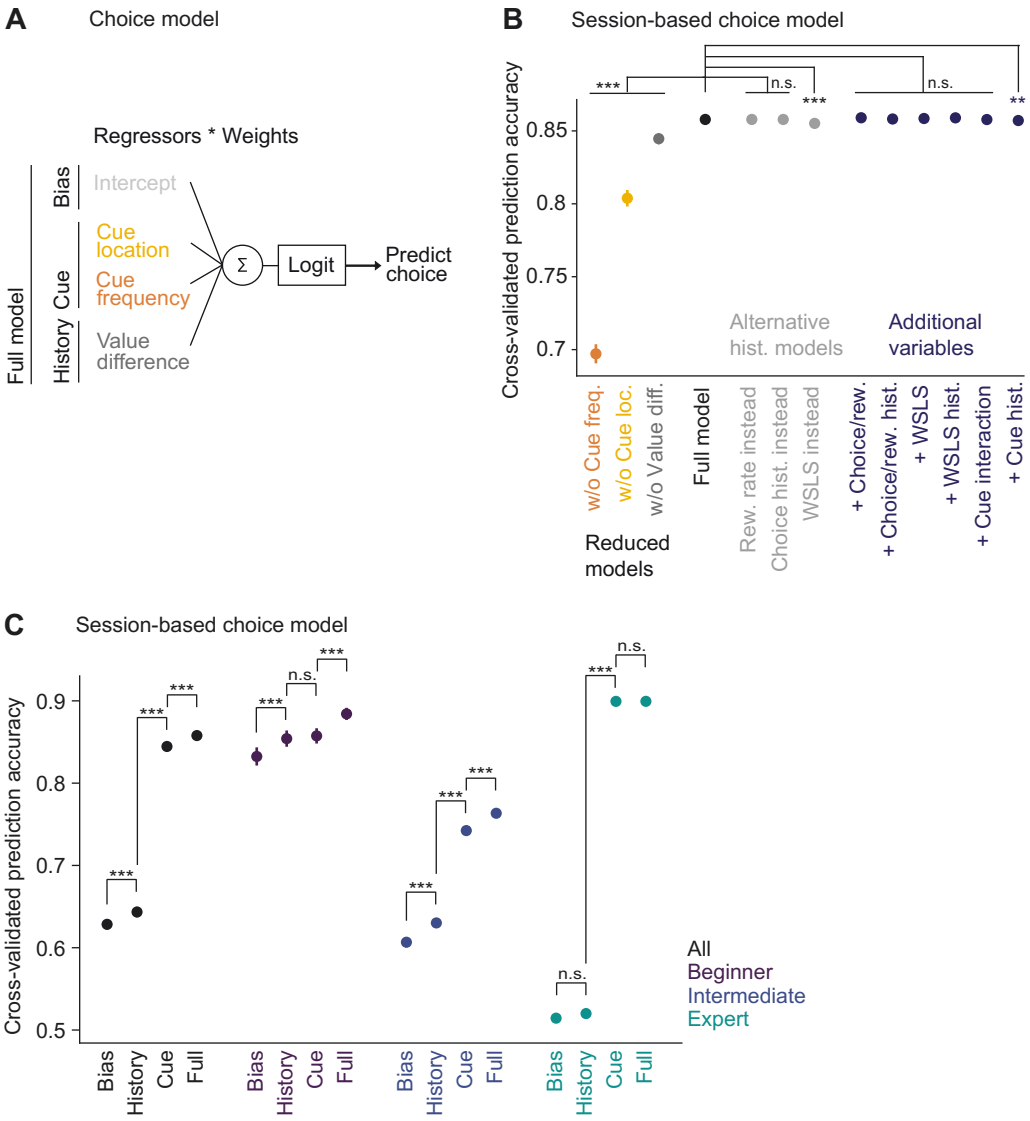


Fig. 8) Session-based choice model
A) Logistic regression model schematic (see Methods for details). **B)** Cross-validated prediction accuracy of the final full model compared to reduced models and alternative models, averaged across sessions. See main text and Methods for details of alternative models. Wilcoxon signed-rank tests with Holm-Bonferroni correction for multiple comparisons across n = 692 sessions. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. **C)** Cross-validated prediction accuracy of the full model compared to reduced models, averaged across sessions, split by performance levels. Wilcoxon signed-rank tests across n = 692 sessions (n = 193 beginner sessions, n = 190 intermediate sessions, n = 309 expert sessions) with Holm-Bonferroni correction for multiple comparisons. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Markers and error bars show mean and standard error of the mean across sessions, respectively. See *Appendix* Table A1 in section 6 for individual statistical test results and sample sizes.

The learning progression can also be examined in the trajectories of predictor weights in the full model across sessions. The weight trajectories are depicted in Figure 9A for an example animal and in Figure 9B for all animals. The weights for the instruction cue location predictor increased during the location task and decreased during the frequency task while they were around zero during the frequency reversed task, indicating that the animals used the location of the cue to guide their choice during the location task and beginning of the frequency task, but not during the frequency reversed task (Fig. 9B, first row). The weights for the instruction cue frequency predictor were around zero during the location task, while they increased towards positive values during the frequency task, and decreased towards negative values in the frequency reversed task, both of which indicated that the animals followed the frequency of the instruction cue to guide the choice (Fig. 9B, second row). The weights for the trial history predictor declined across learning in all three tasks, indicating that the influence of recent history of choices and rewards declined across learning (Fig. 9B, third row). The weights for the intercept (representing a response bias towards one of the two licking spouts) started out with large positive or negative values in the location task, indicating a response bias to the left or to the right, respectively, and gradually approached zero across learning. During the frequency task, the intercept weights remained around zero, while they again started with more extreme values during the frequency reversed task (Fig. 9B, bottom row).

Taken together, both the model comparisons (Fig. 8) and the weight trajectories of the full session-based choice model (Fig. 9) indicated that mice had strong response biases towards one of the two licking spouts in beginner-level sessions, especially in the location task and in the frequency reversed task. Mice further relied on recent history of choices and rewards in beginner-level and intermediate sessions, but not in expert sessions. Finally, mice increasingly followed the instruction cue across learning and the instruction cue was the best predictor of choice in expert sessions.

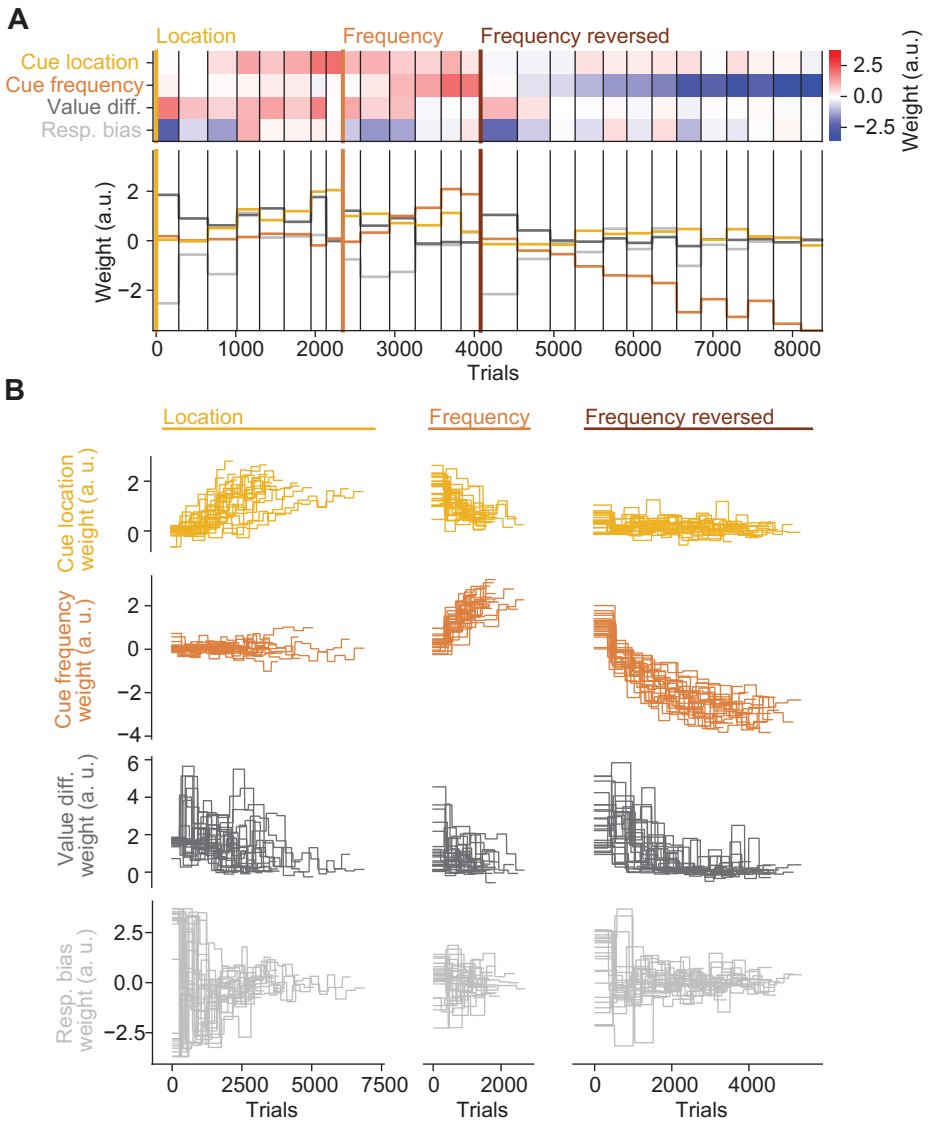


Fig. 9) Weight trajectories from session-based choice model
A) Trajectories of regressor weights across trials for one example animal. Session boundaries are depicted by black vertical lines. Task rule boundaries are depicted by vertical lines color-coded and labeled according to the three task rules. **B**) Trajectories of regressor weights across trials for n = 26 animals, split by the three different task rules and aligned to the rule switches.

3.1.5 Trial-based choice model best explains mouse choices

The session-based choice model revealed relevant predictors of mouse choices during learning, including response biases towards one of the two licking spouts, trial history-related aspects, and instruction cue-related aspects. In intermediate sessions, when the performance was most volatile, the session-based model exhibited the largest fraction of unexplained choices compared to beginner and expert sessions. To test whether the influence of choice predictors, and thus the animals' strategies, fluctuated not only across sessions, but also within sessions, a trial-based model was used. To examine changes across learning at trial-by-trial resolution, the model was first fit by using sliding windows with a subset of trials and thereby increasing the resolution of weight trajectories to a sub-session level (data not shown). This method came at the cost of computational inefficiency and more noisy coefficients due to less data points for fitting. These disadvantages are averted by the PsyTrack model (Roy et al., 2021), which uses hyperparameters to allow predictor weights to fluctuate on a single-trial level. The predictors of the final session-based model were used to fit the PsyTrack model per animal with two different hyperparameters for each predictor that were optimized individually. One of the hyperparameters determined the degree of weight fluctuation across trials, the other one determined the degree of weight fluctuation across sessions (see also *Methods*).

Using the trial-based PsyTrack model, a full model including instruction cue and trial history variables did not show higher prediction accuracy compared to the cue-only model, neither across all sessions, nor for individual performance levels (Fig. 10A). Instead, the trial-based cue-only model performed best among the trial-based models and also performed better than the session-based full model in beginner sessions, intermediate sessions, and sessions following rule switches (Fig. 10B), when performance was volatile (Fig. 6). This indicated that the trial-by-trial fluctuations in response bias, that is, varying weights for the intercept, are sufficient to capture the dependence of the current choice on the recent history of choices. Since these trial-by-trial fluctuations cannot be accounted for in the session-based model, the trial-based model performed better than any session-based model. The cue-only trial-based model also performed best compared to alternative models with other trial history predictors (Fig. 10C), indicating that the fluctuating weights sufficiently captured trial history effects. In expert sessions, the cross-validated prediction accuracy was similar for the trial-based cue-only model compared to the session-based full model (Fig. 10B), which in expert sessions was similar to the session-based cue-only model (Fig. 9B), suggesting that the animals only followed the instruction cues when they had fully learned the task.

The model comparison of session-based models with different trial history predictors had revealed that different representations of the recent history of choices and rewards yielded similar model performance (Fig. 8B). All of these trial history predictors essentially captured a strategy of repeating recent previous choices. In the trial-based model, this strategy is captured by the fluctuating intercept weight, which indicates a response bias towards one of the two licking spouts that is not stable throughout the session (as the intercept in the session-based model), but varies on a trial-by-trial basis. Thus, the trial-based cue-only model with fluctuating weights for intercept and instruction cue predictors captures global response bias strategies and influences of the recent trial history on the current choice, as well as the strategy of following the instruction cue to guide the choice.

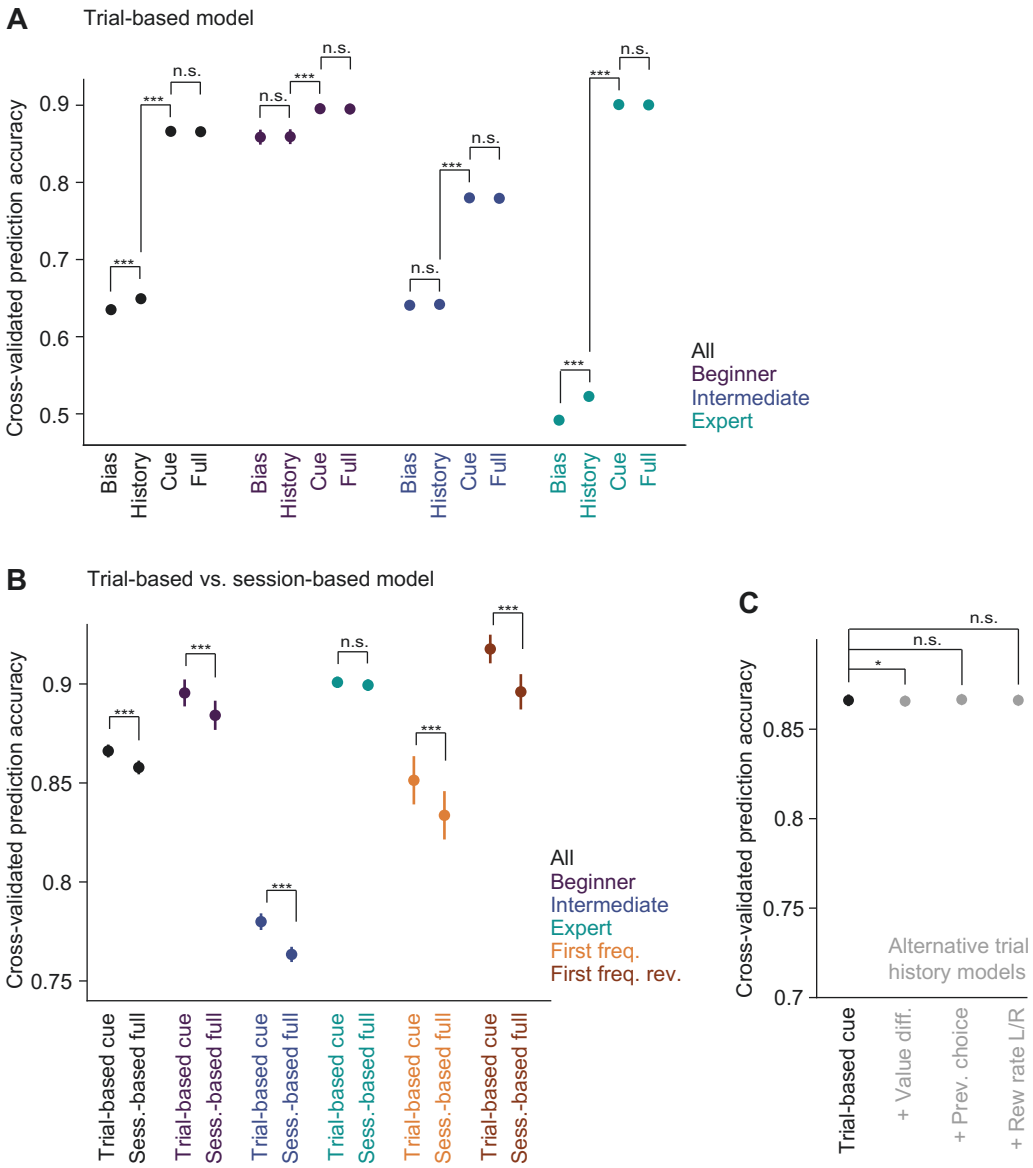


Fig. 10) Trial-based choice model
A) Cross-validated prediction accuracy of the full model compared to reduced models, averaged across sessions, split by performance levels. **B)** Cross-validated prediction accuracy of trial-based vs. session-based models, averaged across sessions, split by performance levels and switch sessions. **C)** Cross-validated prediction accuracy of the final full model compared to alternative models, averaged across sessions. Wilcoxon signed-rank tests across n = 692 sessions (n = 193 sessions, n = 190 intermediate sessions, n = 309 expert sessions) with Holm-Bonferroni correction for multiple comparisons. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Markers and error bars show mean and standard error of the mean across sessions, respectively. See *Appendix* Table A1 in section 6 for individual statistical test results and sample sizes.

A closer look at the weight trajectories of the trial-based cue-only model (Fig. 11B) further elucidates the advantage of trial-by-trial fluctuations of weights over the session-based model. For example, while the session-based model showed a strong response bias weight during the first frequency reversed session (Fig. 9B, bottom row), the trial-based model revealed that the bias weight increased to positive values or decreased to negative values during this session (Fig. 11B, bottom row), indicating a drift towards a leftward or rightward response bias, respectively. Likewise, the cue frequency weight gradually decreased during this session (Fig. 11B, second row). Thus, the trial-based model captured intra-session changes in behavioral strategies, as seen in the analysis in Figure 6.

In summary, both the model comparisons (Fig. 10) and weight trajectories across trials (Fig. 11) of the trial-based choice model revealed an advantage over the session-based model. The trial-based cue-only model reached the highest cross-validated prediction accuracy of all models that were tested, even though it contained the most parsimonious selection of predictors (Fig. 10).

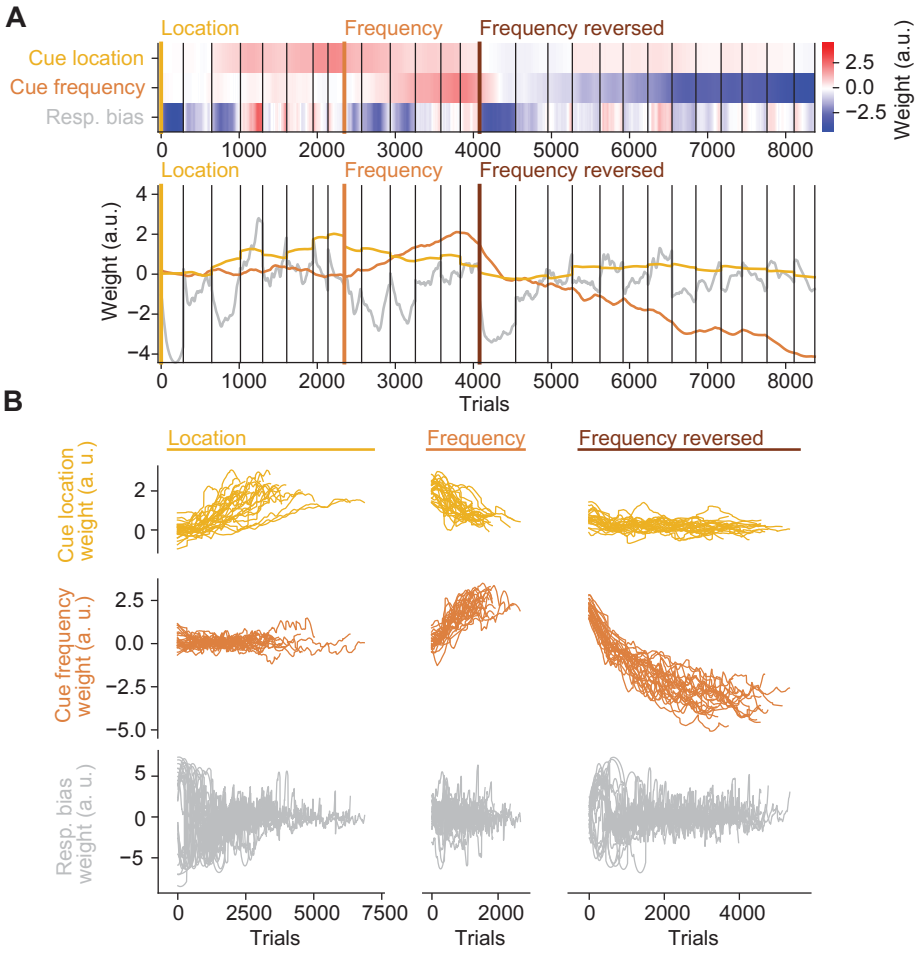


Fig. 11) Weight trajectories from trial-based choice model
A) Trajectories of regressor weights across trials for one example animal. Session boundaries are depicted by black vertical lines. Task rule boundaries are depicted by vertical lines color-coded and labeled according to the three task rules. **B)** Trajectories of regressor weights across trials for n = 26 animals, split by the three different task rules and aligned to the rule switches.

Taken together, the results of both the session-based and trial-based choice models indicated that mice used different strategies across learning to perform the task. In beginner sessions during the location task, when mice learned initial associations, and in beginner sessions during the frequency reversed task, when the contingencies in all conditions were switched compared to the previous task rule, mice showed strong idiosyncratic response biases towards one of the two licking spouts. In beginner sessions during the frequency task, when mice could successfully apply the previous task rule in two out of four conditions, mice did not show strong response biases. In intermediate sessions, the animals relied on trial history-related aspects, which could be captured in the session-based choice model by different representations of repeating previous choices. The trial-based model revealed that fluctuating response bias (i.e., intercept) weights captured these trial history effects and model performance was not improved by additional trial history-related predictors. The trial-based choice model further indicated that the animals' strategies were not static and learning took place also within sessions, since the trial-based model performed better than the session-based model especially in volatile sessions, that is, intermediate sessions and sessions following rule switches. Both session-based and trial-based models (and the comparison of the two) suggested that in expert sessions, mice did not rely on trial history-related aspects, but used only the instruction cues to guide their choice, indicating that they had fully mastered the task.

3.1.6 Sub-trial analysis reveals differences in choice preparation

In order to examine fine-grained changes in behavior at sub-trial resolution, videos were recorded with a camera pointing at the front of the animal to track body part movements during the trial. Example frames of the behavioral videos are depicted in Figure 12A. The left and right hands of the mice were tracked using DeepLabCut (Mathis et al., 2018). Additional body parts such as the nose were also tracked (see *Methods*), but the analysis here is focused on the hand movements during preparation of responses. The relative X positions of the left and right hand are depicted for an example trial in Figure 12B. Analysis of the relative X position of the hands with respect to the video frame (i.e., the horizontal hand position) revealed that compared to the epoch prior to the cue, the animals moved their hands towards the opposite side of the spout that they licked when they made their choice (Fig. 12A/B). Examining the horizontal position of the hands across performance levels revealed that animals changed the horizontal position of the hands, and thus prepared for their response, earlier in expert sessions compared to beginner sessions (Fig. 12C, left). In the location task, both left hand and right hand were moved towards the opposite side of the choice. In beginner sessions, the hands were moved around the time of choice, after the cue epoch of 1000 milliseconds. In expert sessions, the hands were already moved shortly after the onset of the cue, well before the licking response was performed. In the frequency task, the hand movements were more similar in beginner sessions and expert sessions, but still started earlier in expert compared to beginner sessions (Fig. 12C, middle). In the frequency reversed task, the directed hand movements again started earlier in expert compared to beginner sessions (Fig. 12C, right). In beginner sessions in the frequency reversed task, hand positions for left and right choices were crossed during the cue epoch, indicating that the animals first tended towards the wrong side (i.e., followed the previous rule) and then changed the direction.

Overall, the behavior at sub-trial resolution showed that mice prepared for their response during the instruction cue epoch, in expert sessions well before the response lick. In beginner sessions, mice prepared for the response lick later than in expert sessions, especially in the location task and frequency reversed task, when mice showed strong response biases in beginner sessions. The analysis of the sub-trial behavior supports the results presented above that suggest that mice used different strategies during the two rule switches. After the rule switch from location to frequency, mice perseverated on the previous rule and kept responding correctly to conditions in which the required response stayed the same (“stay” trials). Accordingly, the differences between beginner and expert sessions in choice preparation as shown by the hand positions were smaller during the frequency task compared to the other task rules. After the rule switch from frequency to frequency reversed, the animals briefly perseverated on the previous rule and then switched to a response bias strategy (Fig. 6), just as in beginner session during the location task. Accordingly, during the location task and the frequency reversed task, mice showed strong differences between beginner and expert sessions in the choice preparation (Fig. 12C).

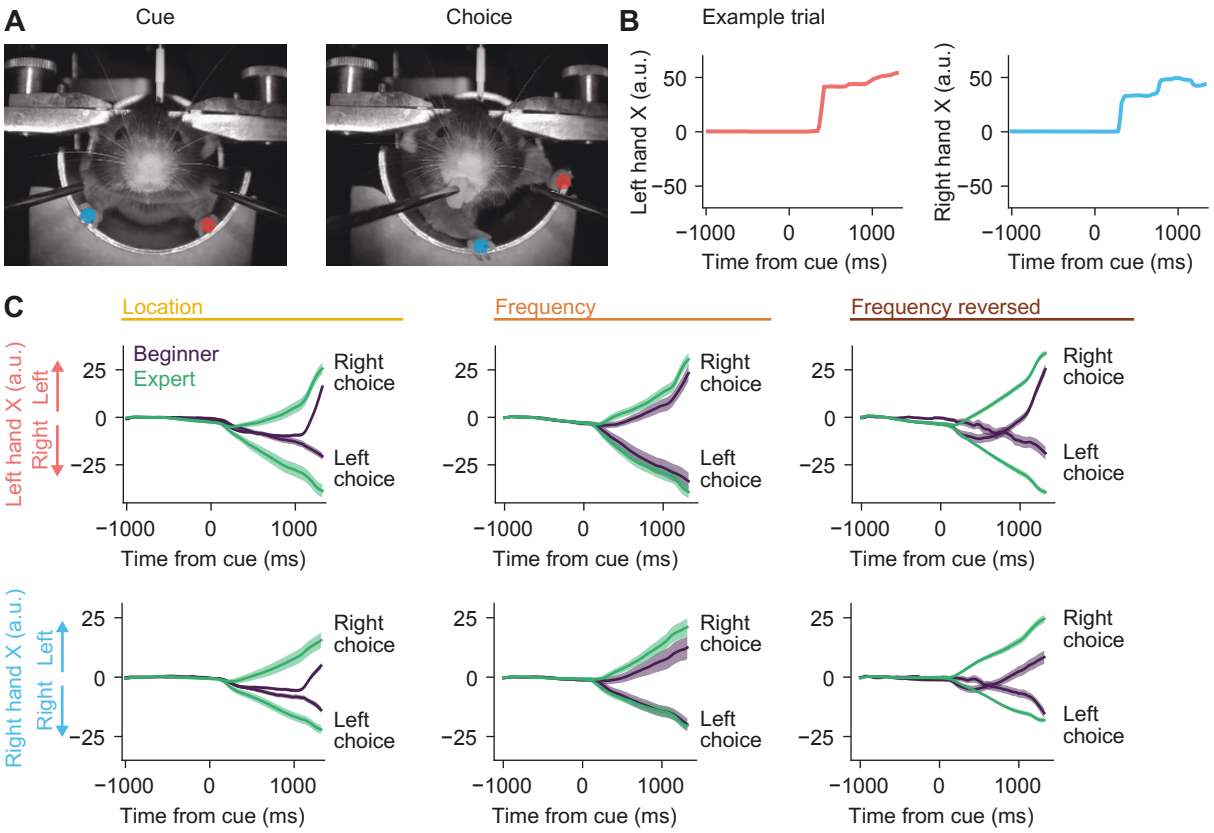


Fig. 12) Behavior at sub-trial resolution
A) Example frames of the behavioral video with body part labels (red dots for the left hand, blue dots for the right hand). Left image was taken at the time of the cue onset, right image at the time of the choice. **B)** Example trial of left-hand and right-hand X positions relative to the trial start. **C)** Average of left-hand and right-hand X positions split by task rules and left-choice vs. right-choice trials for n = 15 animals, correct trials only; lines show means across sessions, shading shows standard error of the mean across sessions.

The results presented so far comprise a multi-timescale behavioral analysis of task acquisition. Mice reliably learned an auditory decision-making task with rule switches over several sessions (Fig. 4). During the initial task rule, mice started out with strong idiosyncratic response biases, licking mainly at one of the two spouts, and gradually learned to follow the instruction cue to guide their choice. Mice responded differently to different rule switches (Fig. 6). After the extra-dimensional rule switch from location to frequency, mice perseverated on the previous rule and kept responding correctly to conditions in which the contingencies stayed that same, while they first responded mostly incorrectly to conditions in which the contingencies had switched. The mice therefore learned the frequency rule without re-entering a response bias strategy. After the intra-dimensional rule switch from frequency to frequency reversed, mice again first perseverated on the previous rule and started the first session after the rule switch with mostly incorrect trials. During this first session, mice reverted to a response bias strategy and licked mostly at one of the two spouts for a few sessions, before they again gradually learned to follow the instruction cue. Choice models (Figs. 8-11) revealed that across learning, in addition to the instruction cue and response bias strategies, mice used the recent history of choices to guide the current choice, but only in beginner and intermediate sessions, not in expert sessions. When mice performed at expert level after they had fully learned the task, they only followed the instruction cues to guide their choices. This suggests that the consolidated behavior was well-controlled and mice made trial-by-trial decisions according to the instruction cue and independent of the history of recent choices and rewards.

3.2 Striatal dopamine signaling during task acquisition

To track dopamine signals at sub-second resolution across the learning process, mice were injected with an adeno-associated virus to express the genetically encoded fluorescent dopamine sensor dLight (Patriarchi et al., 2018) in different subregions of the striatum (see *Methods*). dLight is expressed in the cell membranes of local neurons and changes its fluorescence level upon binding of dopamine, which is released from dopaminergic neurons originating in the midbrain. To image fluorescence changes of the sensor indicating changes in dopamine levels, optic fibers for fiber photometric measurements were chronically implanted above the injection sites (Fig. 13A). Measurements in three subregions of the striatum, the ventral striatum (VS), the dorsomedial striatum (DMS), and the dorsolateral striatum (DLS), were performed separately in three groups of mice (i.e., one subregion per group). Reconstructed implantation locations and example histology images are depicted in Figure 13B. Fiber photometric measurements of dopamine signals were performed in every other behavioral session, resulting in quasi-continuous data points throughout the learning process (Fig. 7).

How dopamine signals developed across learning will be described in the following sections, with a focus on the central questions of how dopamine cue and reward responses evolved across learning (section 3.2.3) and how predictive and reinforcing aspects of reward prediction error signatures were correlated to behavioral signatures of learning (section 3.2.5). Unless otherwise noted, the results are presented in the order of the trial events.

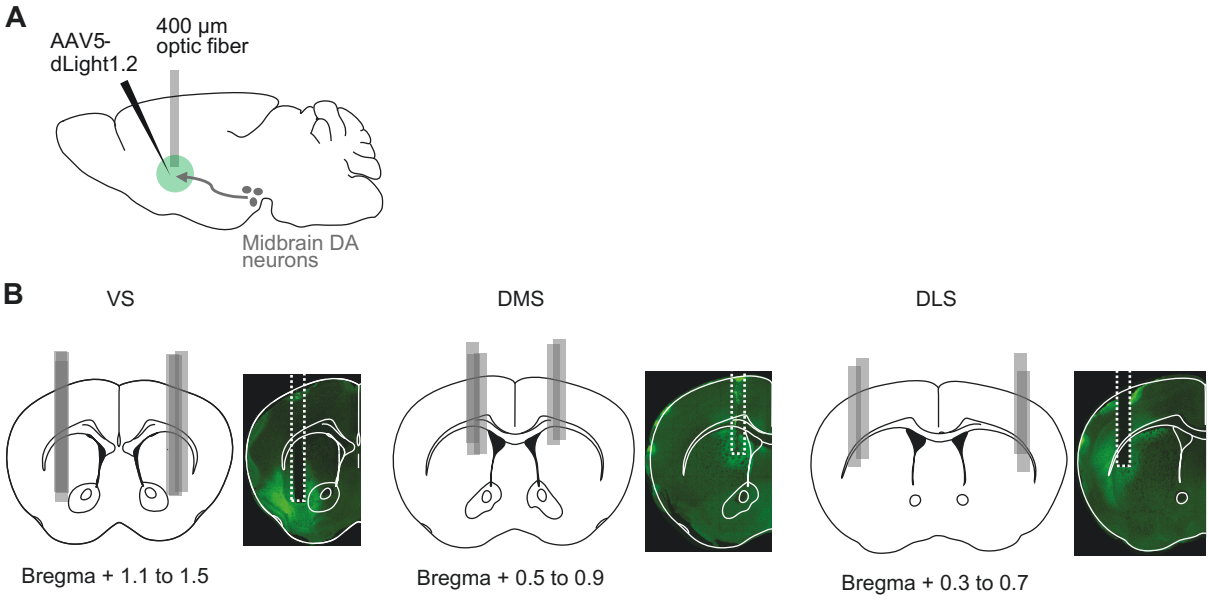


Fig. 13) Implantation locations in fiber photometry experiments
A) Schematic of the virus injection and optic fiber implantation in the striatum (sagittal view). **B)** Approximate implant locations of the optic fibers for all animals (left) and one example animal (right) for three subregions of the striatum: VS (n = 6 animals), DMS (n = 5 animals), and DLS (n = 4 animals). Left: Schematic of coronal sections of the mouse brain with approximate fiber placements (gray bars). Right: Example confocal image of the brain slice with dLight fluorescence, overlayed with mouse brain atlas schematic (solid white lines) and approximate fiber placement (dashed white lines). Mouse brain atlas schematics were adapted from Paxinos and Franklin (2001).

3.2.1 Dopamine signals are modulated across the trial

dLight fluorescence was monitored continuously in each imaging session with stable background fluorescence levels and little photobleaching in all subregions (Fig. 14A). In order to examine dopamine signals in response to different trial events, relative fluorescence values ($\Delta F/F$) were calculated for each trial by subtracting and dividing by the average baseline activity before the trial start (see *Methods* for details). For comparisons across sessions and animals, the $\Delta F/F$ signals were normalized. Figure 14B shows example trials for all three subregions. The trial timing was structured in a way that the only variability within the trial arose from variability in response times of the animals. Figure 14B and all subsequent figures therefore show the trial epochs with two different alignments. The first part of the trial is aligned to the cue and includes the trial start at -1000 milliseconds, the cue at zero milliseconds, and the movement of the spouts and lick response after 1000 milliseconds. The second part of the trial is aligned to the outcome. In correct trials, it includes the reward at zero milliseconds, 2000 milliseconds of reward consumption, and the withdrawal of the spouts after 2000 milliseconds. In false trials, it includes the withdrawal of the spouts at zero milliseconds. For visualization, the outcome period is depicted starting 160 milliseconds after the outcome event to account for latencies in the reward delivery and lags of dLight kinetics as well as to minimize redundancy in visualization between the two alignments.

On average, each trial event led to robust responses in the dopamine signals across all subregions, while instruction cue and outcome elicited the most prominent responses (Fig. 14C). The trial event responses were homogeneous across trials (Fig. 14C, upper panel). The dopamine signals exhibited different dynamics across the three subregions, with the broadest peaks in the VS, narrower peaks in the DMS and the narrowest peaks in the DLS (Fig. 14B/C), a phenomenon that has recently been reported (Wei et al., 2021). In all subregions, the dopamine response to the outcome was strongly modulated by the type of outcome, that is, correct versus false trials. In correct trials, when a reward was obtained, there was a strong positive response in the dopamine signal, while in false trials without a reward, there was a strong negative response (Fig. 14C). This signature is in accordance with the concept of reward prediction errors (Schultz et al., 1997), since the occurrence of a reward elicits a positive reward prediction error, and thus a positive dopamine transient, while the non-occurrence of a reward at the same time point in the trial elicits a negative reward prediction error, and thus a negative dopamine transient.

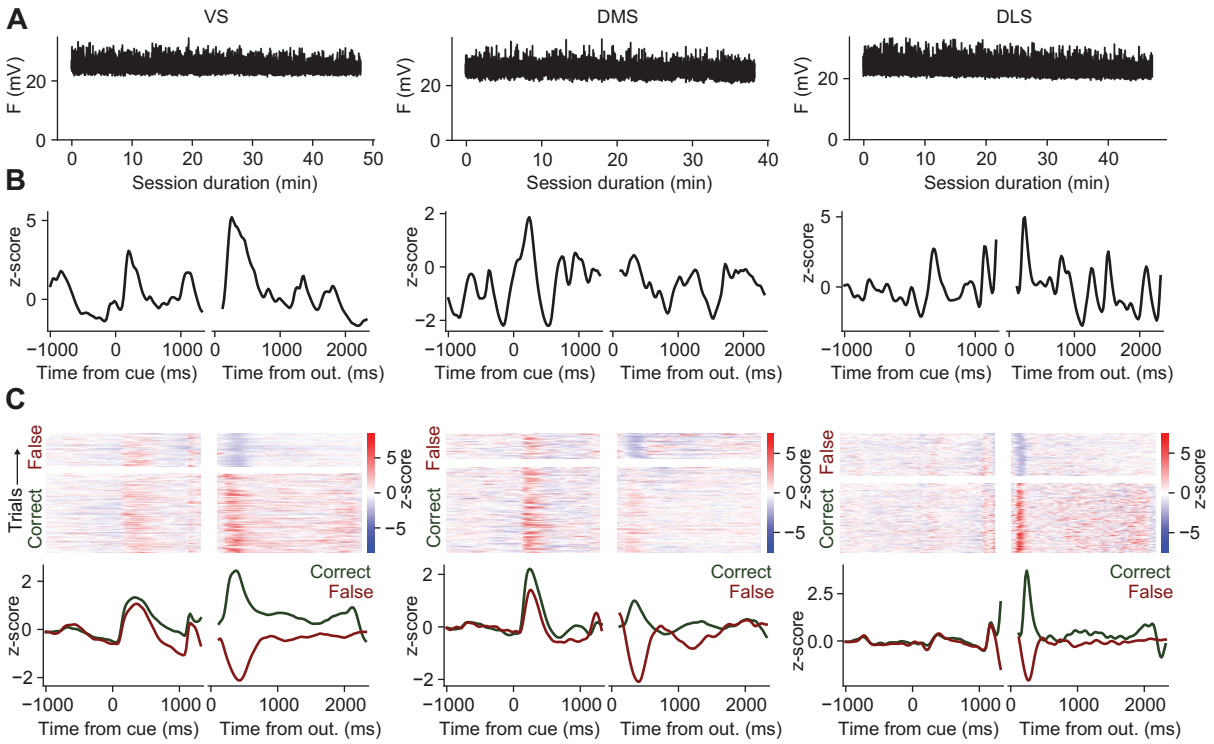


Fig. 14) Example dopamine signals
A) Raw fluorescence traces of example sessions for all subregions. **B)** Normalized fluorescence of example trials for all subregions. Left, aligned to cue; right, aligned to outcome. **C)** Normalized fluorescence in example intermediate sessions for all subregions. Top, fluorescence in each trial, split by correct and false trials; bottom, session average fluorescence, split by correct and false trials.

In order to control for potential movement artifacts in the fluorescence signals due to movements of the brain relative to the fiber, photometric control measurements were performed with a different stimulation wavelength to measure autofluorescence of the tissue independent of changes in dopamine signals (see *Methods*). The assumption was that autofluorescence of the same intensity as the dLight signal should contain similar movement artifacts. In separate sessions before and after task learning, the autofluorescence was measured throughout the whole session. The background fluorescence in these control sessions was adjusted to produce a similar absolute intensity level as in the dLight sessions and showed similarly small degrees of photobleaching across the session (Fig.15A). Comparing the relative fluorescence $\Delta F/F$ averaged across trials and animals revealed that the control signals were orders of magnitude smaller than the dLight signals (Fig. 15B/C) and exhibited different time courses (Fig. 15C). It was therefore concluded that the dLight signals in the head-fixed preparation used in the experiments reported here contained only negligible fluctuations due to movements.

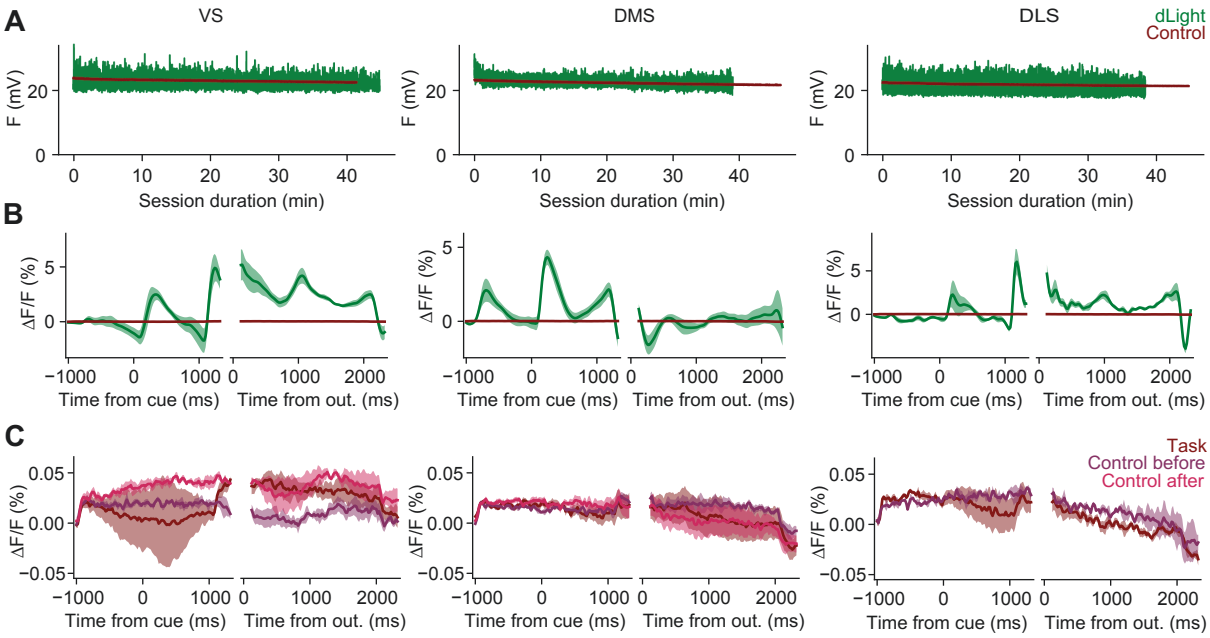


Fig. 15) Movement control measurements
A) Raw fluorescence in an example session with dLight measurement (green) or control measurement (red). **B)** Relative fluorescence for the last task sessions with dLight imaging (green) and a task session after the experiment with autofluorescence control imaging (red); trials are averaged for each session and across animals. Lines show average across animals, shading shows standard error of the mean across animals (VS, n = 6 animals, DMS, n = 5 animals, DLS n = 4 animals). **C)** Same as B) but only control signals from three different sessions before and after the experiment ("Task": last task session, "Control before": pre-training session before task learning, "Control after": pre-training session after task learning). Lines and shading show mean and standard error of the mean across sessions, respectively.

3.2.2 Dopamine cue response builds up during early sessions

To examine changes in dopamine transients during the first sessions in the task, session averages across animals are depicted in Figure 16. Both the time course of dopamine transients (Fig. 16A) and the average amplitude of transients in different trial epochs (Fig. 16B) exhibited changes across the first three task sessions.

During the instruction cue epoch, the dopamine response in the VS and DMS gradually shifted backward in time (Fig. 16A) and increased in amplitude (Fig. 16B) over the first three sessions in the task. During the first task session (*Session 0*), there was an initial small peak at the cue onset, followed by a second peak. Most clearly visible in the DMS, the second peak of the dopamine transient in response to the cue shifted backward in time and moved closer to the cue onset over the first three sessions (Fig. 16A), a phenomenon that has been reported in the VS during Pavlovian learning (Amo et al., 2020). In the DLS, there was little change in the dopamine cue response during the first three sessions. During the spout epoch, that is, when licking spouts were moved into the animals' reach after cue offset (i.e., 1000 milliseconds after cue onset), the dopamine response decreased across the first three sessions in all subregions (Fig. 16A/B). During the outcome epoch, the dopamine response was relatively stable across the first three sessions in all subregions (Fig. 16A/B). Note that for all but two animals the first three sessions were beginner sessions (for two animals the third session was an intermediate-level session), that is, no apparent change in task performance had occurred yet (see also Fig. 4B/D and Fig. 7A for an example).

To better understand the dopamine transients during early sessions in the task, it is helpful to examine pre-training sessions before the task. The animals started the task after two to four sessions of pre-training (see *Methods*), during which no instruction cue was played and only one of the two licking spouts was presented in each trial (with the two spouts randomly alternating across trials). Thus, during pre-training, a reward was obtained in every trial, if the animals licked the presented spout, that is, if they actively collected the reward. The animals collected rewards in virtually every trial during pre-training sessions until they were sated and stopped responding.

Since there was no instruction cue during pre-training, there was also no dopamine cue response in any subregion during pre-training (Fig. 16C). The dopamine spout response increased from pre-training to the first task session in all subregions (Fig. 16C/D). The dopamine spout response also increased between the very first pre-training session (*Pre-training -3*) and the second pre-training session (*Pre-training -2*) in the DMS (Fig. 16C/D middle; dopamine data for the very first pre-training session was only available for DMS). Note that most animals received a further pre-training session between the pre-training session two sessions before the task (*Pre-training -2*) and the first task session (*Session 0*), except for two animals in the DLS group (hence the labels *Pre-training -1/-2* in Fig. 16C right). This may explain why the spout response further increased from pre-training to the first session in the task. The dopamine outcome response increased from pre-training to the first session in the task in all subregions (Fig. 16C/D).

In summary, in the VS and DMS, the dopamine cue response increased and shifted backward in time during the first three sessions in the task. This phenomenon is predicted by reinforcement learning theory (Schultz et al., 1997) and has been previously reported in Pavlovian learning in the VS

(Amo et al., 2020). In the DLS, the dopamine cue response did not change during the first sessions in the task. In all subregions, the spout response increased across pre-training sessions, when animals obtained a reward in close to 100 percent of the trials, and then slowly decreased over the first sessions in the task, when animals performed at beginner level of around 50 percent correct trials and therefore obtained rewards in approximately half of the trials. Thus, the spout movements were certain predictors of the reward during pre-training, while they did not predict rewards during beginner sessions in the task. The increase in dopamine spout responses during pre-training and the decrease during early task sessions is therefore in line with the reward prediction error framework (Schultz et al., 1997). The dopamine reward response increased from pre-training to the first session in the task in all subregions, which was also in accordance with the concept of reward prediction errors, since rewards occurred in virtually all trials during pre-training and therefore were much better predicted than during beginner sessions, when they only occurred in around 50 percent of the trials.

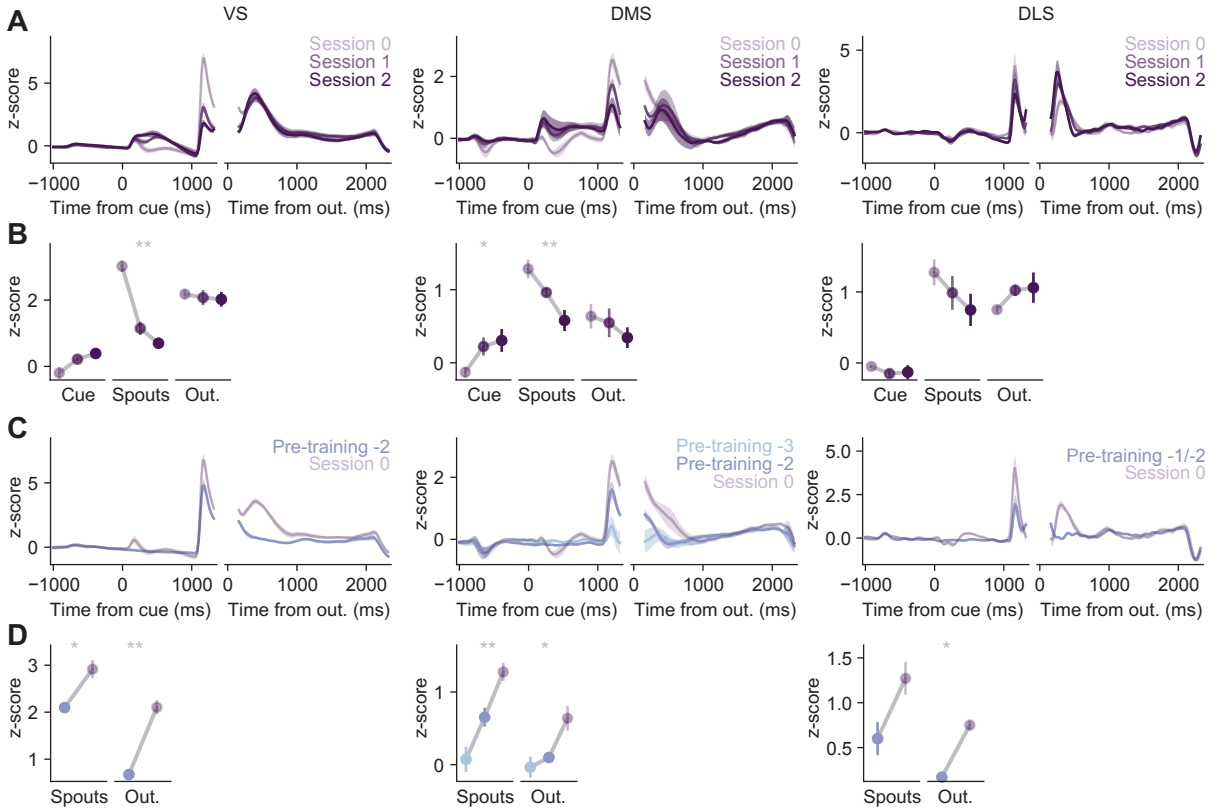


Fig. 16) Dopamine transients in first sessions during initial task rule
A) Average normalized dLight fluorescence in correct trials split by the first three sessions in the task. Lines show mean across sessions, shading shows standard error of the mean across sessions (one session per animal; VS, n = 6 animals, DMS, n = 5 animals, DLS n = 4 animals). **B)** Average normalized dLight fluorescence in correct trials split by the first three sessions in the task, averaged across trial epochs (cue, spouts, outcome); mean ± standard error of the mean across sessions. **C)** Average normalized dLight fluorescence in correct trials split by pre-training and the first session in the task. Lines show mean across sessions, shading shows standard error of the mean across sessions (one session per animal). **D)** Average normalized dLight fluorescence in correct trials, split as in C), averaged across trial epochs (cue, outcome); mean ± standard error of the mean across sessions. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Kruskal-Wallis tests across pooled sessions. See *Appendix* Table A1 in section 6 for individual statistical test results.

3.2.3 Dopamine reward response inversely scales with task performance

A central question in this study was, how dopamine responses to rewards and reward-predicting cues evolve during the acquisition of instrumental associations in the present decision-making task. One hypothesis in accordance with the reward prediction error framework was that cue responses scale with task performance, while reward responses inversely scale with task performance (Hollerman & Schultz, 1998). In order to examine changes of dopamine transients during different trial epochs across the learning process, session averages of dopamine transients were compared across the performance levels “beginner”, “intermediate”, and “expert” (as defined in Fig. 7). Figure 17 shows pooled session averages of dopamine transients throughout the trial across performance levels (beginner, intermediate, expert) and task rules (location, frequency, frequency reversed).

It has been shown above in section 3.2.2 that the dopamine cue response in the VS and DMS shifted backward in time (Fig. 16A) and increased (Fig. 16B) during the first three (beginner) sessions of the task, before increases in task performance. Across all subregions, during subsequent sessions the dopamine cue response overall was relatively stable throughout performance levels and task rules, with a few exceptions (Figs. 17A/B). In the location task the dopamine cue response slightly increased across performance levels, with the clearest increase in the DMS (Fig. 17A first row, Fig. 17B). During the frequency task, the dopamine cue response was relatively stable (Fig. 17A second row, Fig. 17B). During the frequency reversed task, there was an increase in the dopamine cue response in beginner sessions compared to intermediate and expert sessions both in the VS and in the DMS. These differences in dopamine cue responses after rule switches were further examined by taking a closer look at sessions following rule switches and distinct trial types, as reported below (sections 3.2.4 and 3.2.5). Overall, the dopamine cue response was not simply scaled to task performance throughout the task rules, but instead exhibited a more complex pattern, which will be examined in detail in the following sections.

The dopamine response during the spout epoch after cue offset (i.e., 1000 milliseconds after cue onset) was shown to decrease during the first (beginner) sessions in the task in Figure 16. During the first task rule (location task), in all subregions, the dopamine spout response decreased from beginner level to intermediate level of performance and subsequently stayed at a constant level across the other task rules (Fig. 17A/C; the only exception of an increased dopamine spout response in the DMS in beginner sessions during the frequency reversed task is a co-product of the increased and prolonged dopamine cue response during these sessions). Overall, Figure 17 supports the notion presented above (section 3.2.2 and Fig. 16) that the dopamine spout response decreased during early sessions of the task, in which the spouts were no longer a certain predictor of reward (in contrast to pre-training). The dopamine spout response subsequently played a minor role during learning with little changes across performance levels.

The dopamine response during the outcome epoch in correct trials (i.e., the dopamine reward response) exhibited the most straightforward changes across performance levels compared to the other dopamine transients in the trial. In all subregions and across all task rules, the dopamine reward response declined from beginner to intermediate to expert sessions (Fig. 17A/D). Thus, the dopamine reward response was inversely scaled to task performance. After rule switches, the dopamine reward

response increased and then decreased again across learning (Fig. 17A/D). The dopamine reward response was therefore reset together with behavioral performance after rule switches, which suggests that the dopamine reward response rapidly tracked changes in contingencies and average reward expectations. This pattern was in accordance with the reward prediction error framework (Schultz et al., 1997), since rewards were better predicted at high performance levels, and in line with a previous report (Hollerman & Schultz, 1998). In addition, the dopamine outcome response in false trials without rewards was consistently negative across all task rules and subregions (*Appendix* Fig. A1), indicating negative reward prediction errors in line with the theoretical framework (Schultz et al., 1997). Negative reward prediction errors were not scaled to performance.

In summary, across all subregions and task rules, in accordance with the reward prediction error framework, the dopamine reward response was inversely scaled to task performance and thus average reward expectation. The session averages suggested that the dopamine reward response rapidly increased after rule switches. The time course of this increase of dopamine reward responses after rule switches will be examined at closer detail in the next section. In contrast to the clear pattern of the dopamine reward response, the dopamine cue response was not simply scaled to performance, especially after rule switches. Instead, the dopamine cue response exhibited more complex changes, which will be examined more closely in sections 3.2.4 and 3.2.5.

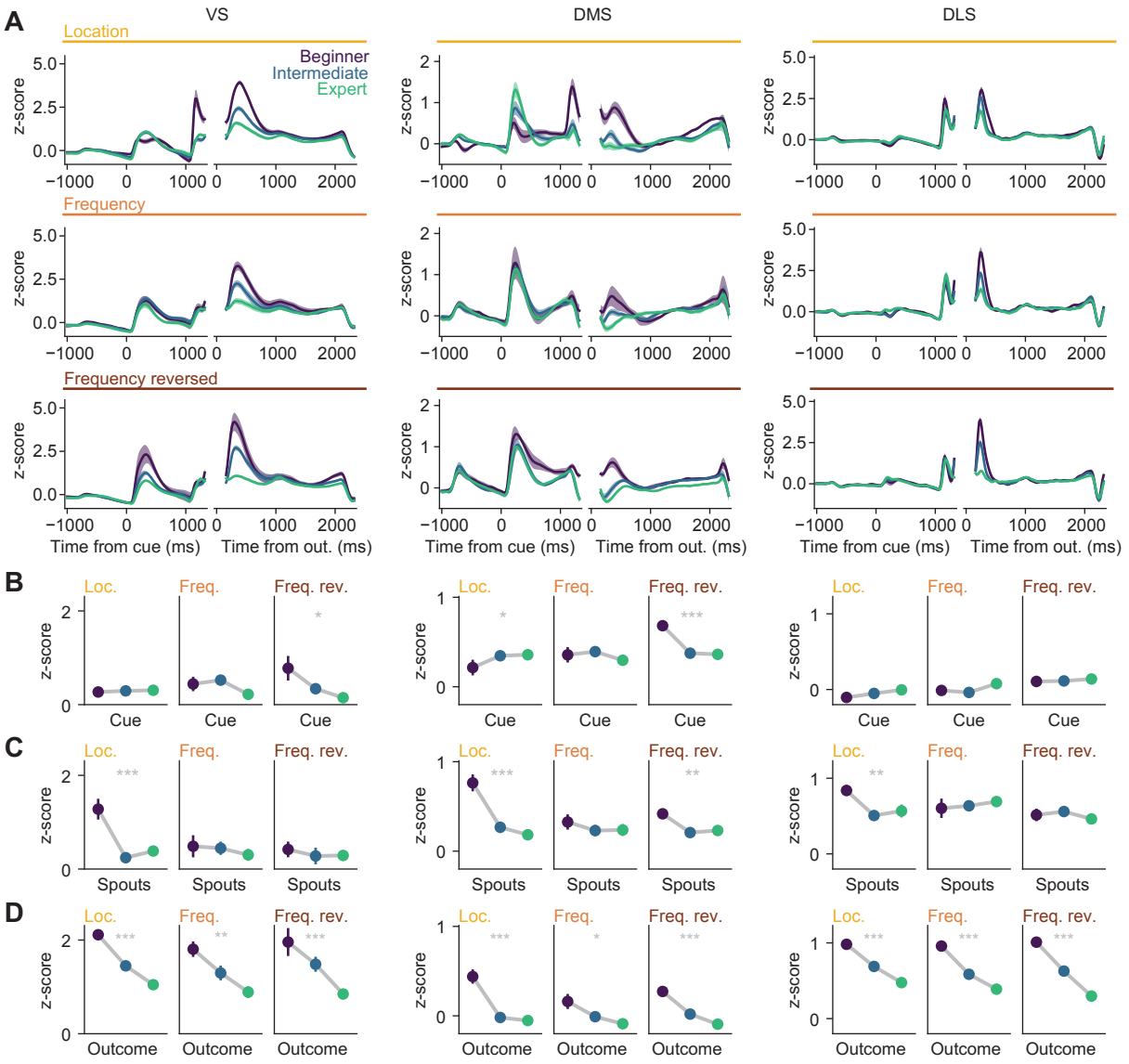


Fig. 17) Dopamine transients across performance levels and task rules
A) Average normalized dLight fluorescence in correct trials split by performance levels (beginner, intermediate, expert) for all task rules (location, frequency, frequency reversed). Lines show mean across sessions, shading shows standard error of the mean across sessions (VS, n = 6 animals, DMS, n = 5 animals, DLS n = 4 animals, see *Appendix* Table A1 in section 6 for number of sessions per group). **B-D)** Average normalized dLight fluorescence in correct trials, split as in A), averaged across trial epochs (cue, spouts, outcome); mean ± standard error of the mean across sessions. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Kruskal-Wallis tests across pooled sessions. See *Appendix* Fig. A1 in section 6 for a visualization of A) in error trials. See *Appendix* Fig. A2 in section 6 for a visualization of A) using mean and standard error of the mean across animals; see *Appendix* Fig. A3 in section 6 for a visualization of A) for each animal separately; see *Appendix* Table A1 in section 6 for number of sessions per group and individual statistical test results.

3.2.4 Dopamine reward response quickly adapts after rule switches

The analysis of dopamine transients across performance levels showed that the dopamine reward response increased after rule switches (Fig. 17A/D), which is in accordance with the concept of reward prediction errors, since the performance and therefore the reward expectation dropped after rule switches. To further examine the time course of this increase in dopamine reward response, the sessions surrounding rule switches (i.e., last session in the previous task and first session in the current task) are depicted in Figure 18. Here, the dopamine reward response will be described first, followed by the dopamine cue response.

In all three subregions, the dopamine reward response was increased in the first frequency session compared to the last location session (Fig. 18A/B). Likewise, in all three subregions the dopamine reward response was increased in the first frequency reversed session compared to the last frequency session (Fig. 18C/D). To examine how fast this increase occurred, in addition to the average dopamine transients of correct trials in the last session before the rule switch and the first session after the rule switch, Figure 18 shows the average dopamine transients of the first 10 percent of correct trials in the first session after the rule switch. In all subregions and after both rule switches, the dopamine reward response was already increased during the first 10 percent of correct trials after the rule switch (Fig. 18A-D). In the VS and DLS, the dopamine reward response was relatively stable across the session after this fast initial increase. In the DMS, the dopamine reward response decreased again after the initial increase, especially after the rule switch from frequency to frequency reversed (Fig. 18D, middle). Overall, the dopamine reward response quickly adapted to changes in contingencies after rule switches, suggesting that the dopamine reward response faithfully and accurately tracked reward expectations.

In contrast to the straightforward pattern of changes in the dopamine reward response after rule switches, the dopamine cue response surrounding rule switches was more complex. The pattern of the average dopamine cue response after rule switches was a result of both the distinct behavioral changes associated with the two rule switches and the modulation of the dopamine cue response by behavioral preference for particular conditions, which will be described in detail in section 3.2.5. In the VS and DMS, after the rule switch from location to frequency, the dopamine cue response first increased and then decreased again (Fig. 18A/B). After the rule switch from frequency to frequency reversed, the dopamine cue response was not increased immediately during the first 10 percent of trials but increased throughout the rest of the session (Fig. 18C/D). In the DLS, there was little change in the dopamine cue response in the VS and DMS after rule switches (Fig. 18A-D). The rapid temporary increase of the average dopamine cue response in correct trials after the first rule switch and the overall increase of the average dopamine cue response in correct trials after the second rule switch in the VS and DMS were a result of the distinct changes in the composition of correct trials and false trials in preferred and non-preferred conditions across the sessions following the two different rule switches. This pattern will be described again in detail at the end of section 3.2.5 after a description of the modulation of dopamine cue and reward responses by preference for particular conditions.

In summary, the dopamine reward response quickly adapted to changes in contingencies after rule switches and tracked changes in average reward rate and corresponding reward expectations. This is in accordance with the reward prediction error framework, since reward prediction errors become smaller when rewards are well predicted during expert sessions before rule switches and become larger when rewards are less well predicted during beginner sessions after rule switches.

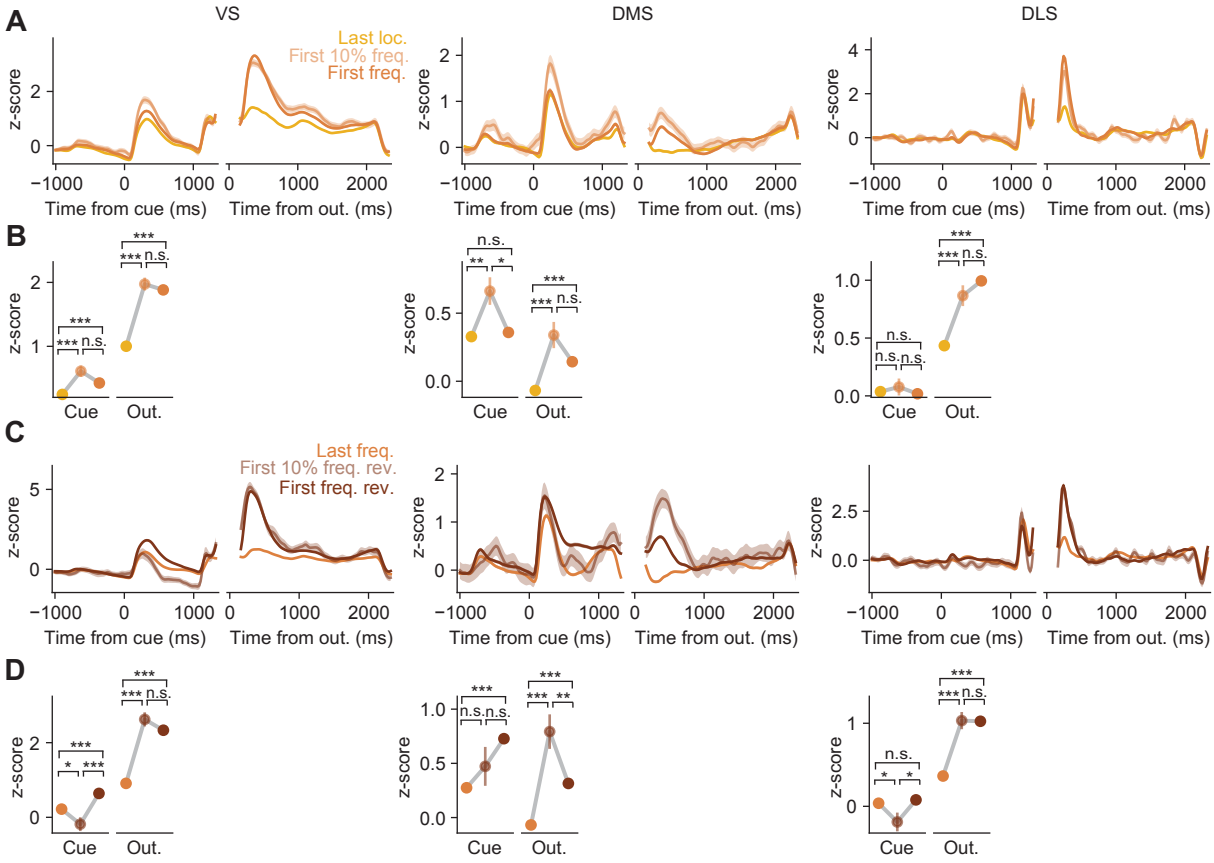


Fig. 18) Dopamine transients during rule switches
A) Average normalized dLight fluorescence in correct trials split by last location session, first frequency session (first 10 percent of trials), and first frequency session (whole session). Lines show mean across trials, shadings show standard error of the mean across trials (VS, n = 6 animals; DMS, n = 5 animals; DLS, n = 4 animals; see *Appendix* Table A1 in section 6 for number of trials per group). **B)** Average normalized dLight fluorescence in correct trials, split as in A), averaged across trial epochs (cue, outcome); mean ± standard error of the mean across trials. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Wilcoxon rank-sum tests across pooled trials with Bonferroni correction for multiple comparisons. See *Appendix* Table A1 in section 6 for number of trials per group and individual statistical test results. **C)** Same as A), but for rule switch from frequency to frequency reversed. **D)** Same as B), but for rule switch from frequency to frequency reversed.

3.2.5 Partial lag of reward prediction error signature relative to behavior

Another central question in the present study was, how predictive and reinforcing properties of dopamine reward prediction error signatures contribute to instrumental learning (Coddington & Dudman, 2018; Pan et al., 2021). In order to understand the pattern of dopamine transients during cue and reward epoch across learning and how this pattern corresponds to the concept of reward prediction errors, preferred and non-preferred conditions were examined. The behavioral analysis had revealed that after the rule switch from location to frequency mice perseverated and kept applying the location rule in the beginning of the first frequency session (Fig. 6), that is, mice kept responding correctly to the conditions in which the rule stayed the same (preferred or “stay” conditions) and responded mostly incorrectly to the conditions in which the rule had switched (non-preferred or “switch” conditions). To understand how the dopamine signature changed in response to the rule switch, Figure 19 shows dopamine transients split by preferred (stay) conditions and non-preferred (switch) conditions for the sessions surrounding the rule switch from location to frequency.

The results will first be described for the VS. There was no difference between the dopamine transients in stay and switch conditions during the last location task session (*L-1*), when this distinction was not yet relevant (Fig. 19A first row). During the first session of the frequency task (*F0*), the dopamine cue response was larger in stay conditions compared to switch conditions, while there was no difference in the dopamine reward response (Fig. 19A second row). The dopamine cue response difference was maintained during the second and third session in the frequency task (*F1, F2*). In these sessions, there was also a difference in the dopamine reward response between stay and switch conditions, but of opposite direction compared to the dopamine cue response (Fig. 19A third and fourth row). Notably, the behavioral difference between stay and switch conditions (i.e., difference in fraction of correct trials) was largest in the first session after the rule switch (Fig. 19B right-most subpanel), during which there was only a difference in dopamine cue response, but not the dopamine reward response (Fig. 19B left subpanel). The pattern of dopamine responses during cue and outcome epochs in the sessions surrounding the switch from location to frequency is summarized in Figure 19C, which shows the difference between average dLight fluorescence in preferred (stay) compared to non-preferred (switch) conditions. Figure 19D shows the corresponding difference in behavioral performance. The pattern seen in the VS was qualitatively similar in the DMS, although the magnitude of the effect was smaller. In the DLS, there was only a small difference between stay and switch conditions in the dopamine reward response two sessions after the switch. Note that the behavioral pattern was very similar across the three subregions (Fig. 19D).

In summary, in the VS (and to a smaller degree also in the DMS), the dopamine cue response adapted quickly to the switch in contingencies in correspondence to behavioral preferences (Fig. 19D) and accompanying reward expectations. After the rule switch, the dopamine cue response increased for preferred conditions in which a reward was expected, and decreased in non-preferred conditions in which no reward was expected (Fig. 19B/C). In contrast, the dopamine reward response lagged behind the behavioral pattern by at least one session, that is, it was increased in non-preferred conditions and decreased in preferred conditions only two sessions after the rule switch, while the strongest difference in behavior was in the first session after the rule switch (Fig. 19B-D).

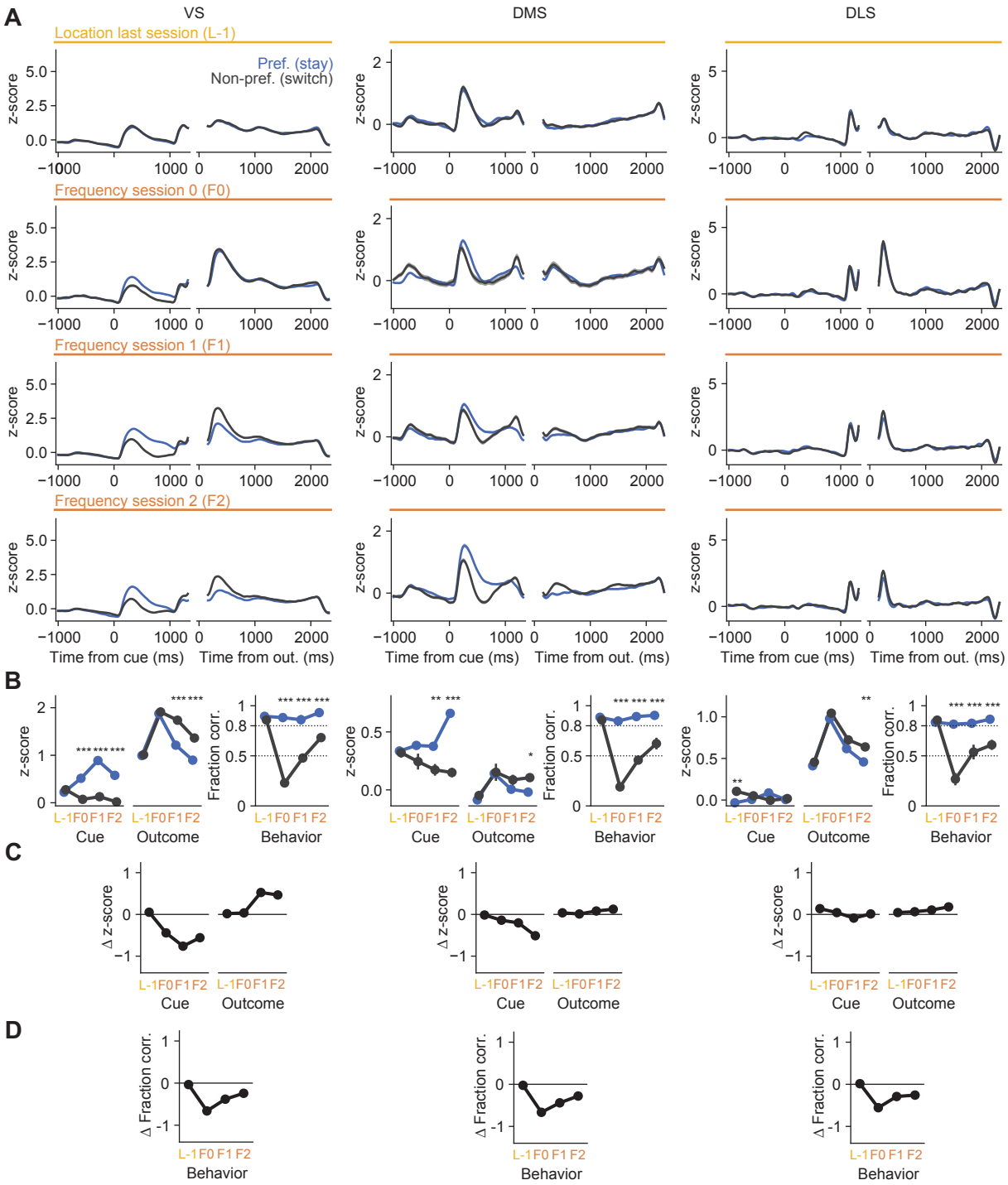


Fig. 19) Dopamine transients in preferred (stay) vs. non-preferred (switch) trials before and after rule switch from location to frequency

A) Average normalized dLight fluorescence in correct trials split by preferred (stay) trials vs. non-preferred (switch) trials for the last session in the location task (L-1) and the first three sessions in the frequency task (F0, F1, F2). Lines show mean across trials, shadings show standard error of the mean across trials (VS, n = 6 animals; DMS, n = 5 animals; DLS, n = 4 animals; see *Appendix* Table A1 in section 6 for number of trials per group). **B)** Average normalized dLight fluorescence, split as in A), averaged across trial epochs (cue, outcome); Wilcoxon rank-sum tests across pooled trials with Bonferroni correction for multiple comparisons. Right-most panel: Average fraction of correct trials, split as in left panel; Chi-squared tests. *** p < 0.001, ** p < 0.01, * p < 0.05. See *Appendix* Table A1 in section 6 for number of trials per group and individual statistical test results. **C)** Difference between average normalized dLight fluorescence (Δ z-score) in preferred (stay) vs. non-preferred trials (switch) trials, split as in A) and B). **D)** Difference between average fraction of correct trials (Δ Fraction correct) in preferred (stay) vs. non-preferred trials (switch) trials, split as in A) and B).

Examining the pattern seen in Figure 19 across performance levels during the whole frequency task (Fig. 20) confirmed the effect seen during the first three sessions. The dopamine cue response was increased in preferred (stay) conditions and decreased in non-preferred (switch) conditions already in beginner sessions and throughout all performance levels (Fig. 20B/C). The dopamine reward response was increased in non-preferred (switch) conditions and decreased in preferred (stay) conditions in intermediate and expert sessions, but not in beginner sessions (Fig. 20B/C). The difference in dopamine transients and corresponding behavioral difference are again summarized in Figure 20C and Figure 20D, respectively. Notably, in all subregions the strongest difference between preferred (stay) and non-preferred (switch) conditions in dopamine responses was in intermediate sessions (Fig. 20C), while the strongest behavioral difference was in beginner sessions (Fig. 20D). Furthermore, the difference in dopamine responses persisted in expert sessions, when the behavioral difference was small (yet significant) and animals performed in both preferred and non-preferred conditions at a level of above 0.8 fraction correct (Fig. 20B). The observed pattern was strongest in the VS, but also observed to a smaller degree in the DMS (Fig. 20). Note that the behavioral pattern was again very similar across the three subregions (Fig. 20D).

In summary, after the rule switch from location to frequency, when mice preferred stay conditions over switch conditions (as observed in their behavioral performance), the dopamine signature of a relative difference between preferred (stay) and non-preferred (switch) conditions was in accordance with the reward prediction error framework. The instruction cue elicited a larger dopamine response in preferred compared to non-preferred conditions, while the reward elicited a larger dopamine response in non-preferred over preferred conditions. This matches reward prediction errors, since the reward expectation is higher and the reward is better predicted in preferred compared to non-preferred conditions. Interestingly, this dopamine signature of a relative difference of the cue and reward response in preferred versus non-preferred conditions partly lagged behind the behavioral signature. Specifically, the dopamine reward response was initially at a similar level for the two conditions, in which the animals showed very different levels of behavioral performance. Only over the course of a few sessions, a dopamine signature corresponding to the behavior emerged and persisted throughout expert sessions, during which there was only a small behavioral difference. Overall, the dopamine signature of dopamine cue and reward responses matched reward prediction errors in intermediate and expert sessions, but not in beginner sessions. Thus, the reward prediction error signature partially lagged behind the behavioral signature and was corrupted during beginner sessions after the rule switch. The reward prediction error signature was intact again, when the behavioral performance increased and mice performed at intermediate and expert level. This effect was observed most clearly in the VS, to a smaller degree in the DMS, but not in the DLS, where there was only weak modulation of the dopamine cue response by preferred and non-preferred conditions.

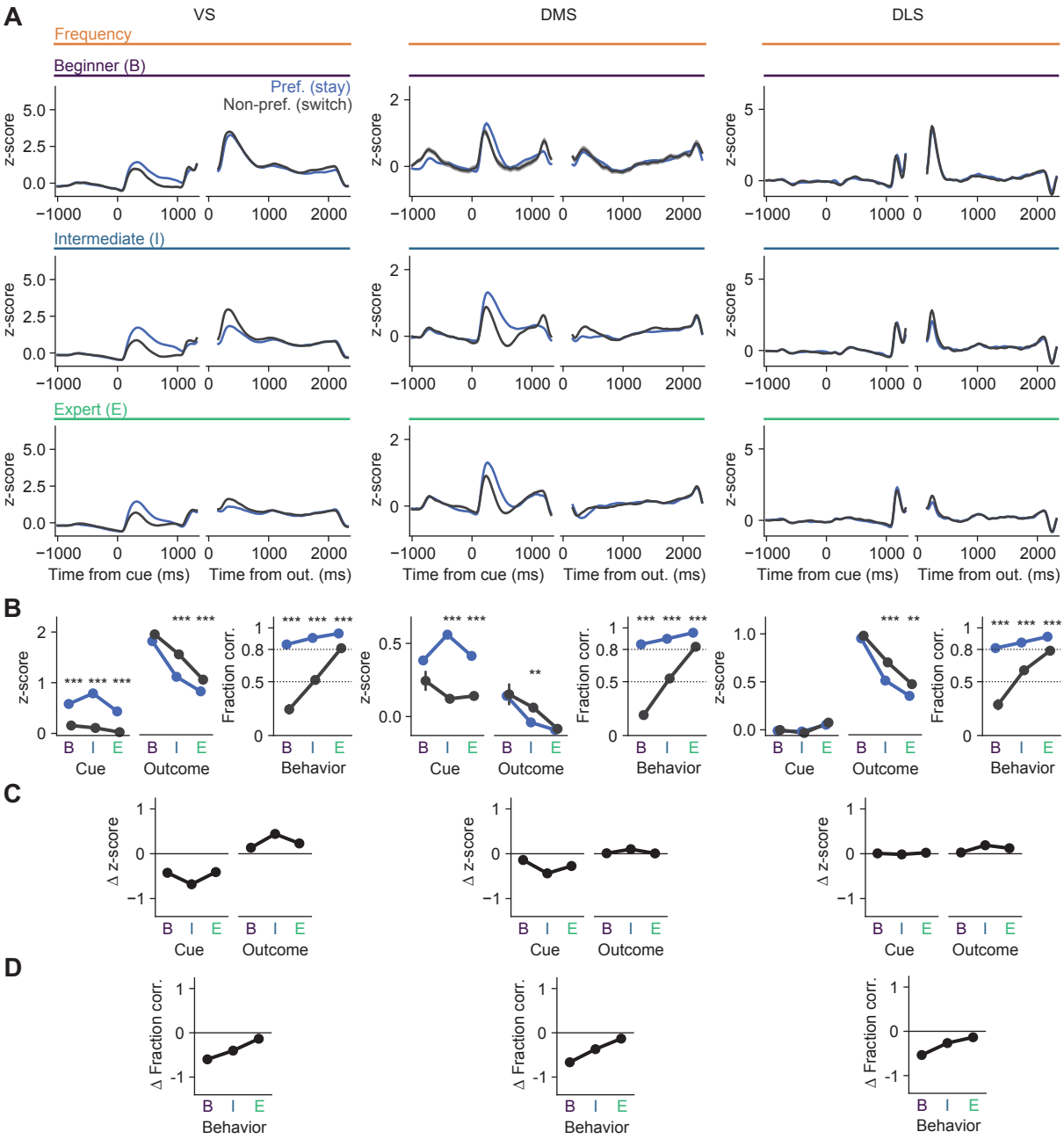


Fig. 20) Dopamine transients in preferred (stay) vs. non-preferred (switch) trials across performance levels in the frequency task

A) Average normalized dLight fluorescence in correct trials split by preferred (stay) trials vs. non-preferred (switch) trials across performance levels in the frequency task. Lines show mean across trials, shadings show standard error of the mean across trials (VS, n = 6 animals; DMS, n = 5 animals; DLS, n = 4 animals; see *Appendix* Table A1 in section 6 for number of trials per group). **B)** Average normalized dLight fluorescence, split as in A), averaged across trial epochs (cue, outcome); mean ± standard error of the mean across trials, Wilcoxon rank-sum tests across pooled trials with Bonferroni correction for multiple comparisons. Right-most panel: Average fraction of correct trials, split as in left panel; Chi-squared tests. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. See *Appendix* Table A1 in section 6 for number of trials per group and individual statistical test results. **C)** Difference between average normalized dLight fluorescence in preferred (stay) vs. non-preferred trials (switch) trials (Δ z-score), split as in A) and B). **D)** Difference between average fraction of correct trials in preferred (stay) vs. non-preferred trials (switch) trials (Δ Fraction correct), split as in A) and B).

The behavioral preference for stay over switch conditions after the rule switch from location to frequency was a consequence of the experimental manipulation (i.e., the rule switch). The contingencies stayed the same for stay conditions and the animals could successfully apply the previous rule, while the contingencies changed for switch conditions and the animals had to update the association between the instruction cue and the required response to obtain the desired outcome. To examine whether the reward prediction error signature and the partial lag relative to behavior observed during the frequency task also generalized during the other task rules, a different definition of preference was necessary, since there was no experimentally induced preference for certain conditions during the other tasks. During the location task and the frequency reversed task, mice exhibited strong response biases in beginner sessions (e.g., Fig. 4) and thus showed a self-selected preference for certain conditions depending on the side of the correct response. To match the timescale of the emergence of reward prediction error signatures over several sessions during the frequency task, a global response preference was chosen for the analogous analysis in the location task and frequency reversed task. This global response preference was defined for each animal as the overall preferred response side across all imaging sessions within the current task.

It has to be noted that in contrast to the experimentally induced preference for stay over switch trials, the individual self-selected global preference was more variable within animals. While most animals consistently preferred one response side over the other during a given task rule or even throughout the whole experiment, they could also vary their preference between sessions. Moreover, in contrast to the experimentally induced preference for stay over switch trials, which was confined to the sessions following the rule switch from location to frequency, animals exhibited self-selected side preferences throughout the whole experiment and thus already had a history of side preference before the rule switch from frequency to frequency reversed, which influenced the starting condition for the preference-related reward prediction error signatures. Even though the animals performed at expert level with very small side biases before the rule switch, the reward prediction error signatures have been shown above to persist throughout expert sessions in the absence of strong behavioral preferences (Fig. 20). Therefore, the effects of slowly emerging reward prediction error signatures may be less clearly observable for the self-selected global side preference compared to the preference for stay over switch trials.

In the VS, after the rule switch from frequency to frequency reversed, a similar effect was observed for preference-related dopamine signatures compared to the effect seen for the preference for stay over switch trials during the frequency task (Fig. 20). In beginner sessions, the dopamine cue response in the VS was larger for preferred compared to non-preferred conditions, while the dopamine reward response was at the same amplitude (Fig. 21A/B). In intermediate sessions, the dopamine cue response was larger in preferred compared to non-preferred conditions, while the dopamine reward response was larger for non-preferred compared to preferred conditions (Fig. 21A/B), which was the expected pattern according to the reward prediction error framework. While the difference in dopamine cue response between preferred and non-preferred conditions was tracking the behavioral difference well, the difference in the dopamine reward response was lagging behind a few sessions. Therefore, the signature of reward prediction errors (Fig. 21C) was again not co-occurring with the largest

difference in behavior (Fig. 21D), but lagging behind. In the DMS, during the frequency reversed task dopamine transients in response to preferred versus non-preferred conditions were modulated in a way consistent with reward prediction errors, with dopamine cue responses larger in preferred trials compared to non-preferred conditions and dopamine reward responses larger in non-preferred compared to preferred conditions (Fig. 21A/B). This pattern was observed across all performance levels with no apparent lags behind behavior (Fig. 21C/D). In contrast to the VS, where there was a very strong difference in the cue response between preferred and non-preferred conditions during beginner sessions, this difference was relatively small in the DMS. Note that also the behavioral difference in beginner sessions was stronger in the VS compared to the DMS (Fig. 21B), indicating that VS animals showed a more consistent global side preference during these sessions. In the DLS, the dopamine cue response was larger in preferred conditions compared to non-preferred conditions during intermediate and expert sessions, while there was no modulation of the dopamine reward response (Fig. 21A-D right column).

Not only after rule switches, but also during the initial acquisition of the location task rule, the VS dopamine transients exhibited a similar pattern in response to behavioral preference as described above, with larger dopamine cue responses for preferred compared to non-preferred conditions and larger dopamine cue responses in non-preferred compared to preferred conditions (Fig 22A/B). In the VS, the largest difference in the dopamine reward response again lagged behind the largest behavioral difference (Fig. 22C/D). Note that during beginner sessions the dopamine cue response was not yet modulated (Fig. 22A/B). In contrast to the other task rules, the animals started the location task from a naïve state with no prior preferences or knowledge of the conditions, which may explain the slower emergence of a preference modulation of the dopamine cue response. In the DMS, the dopamine cue response was again modulated by global response preference in line with reward prediction errors, while the dopamine reward response slightly deviated from this pattern during intermediate and expert sessions (Fig. 22C). Note that in the DMS during the location task the global response preference was reversed during intermediate sessions (Fig. 22D), which may explain this deviation. In the DLS, there were only small modulations by global response preference during the location task, albeit also exhibiting a pattern in line with reward prediction errors (Fig. 22A-D right column)

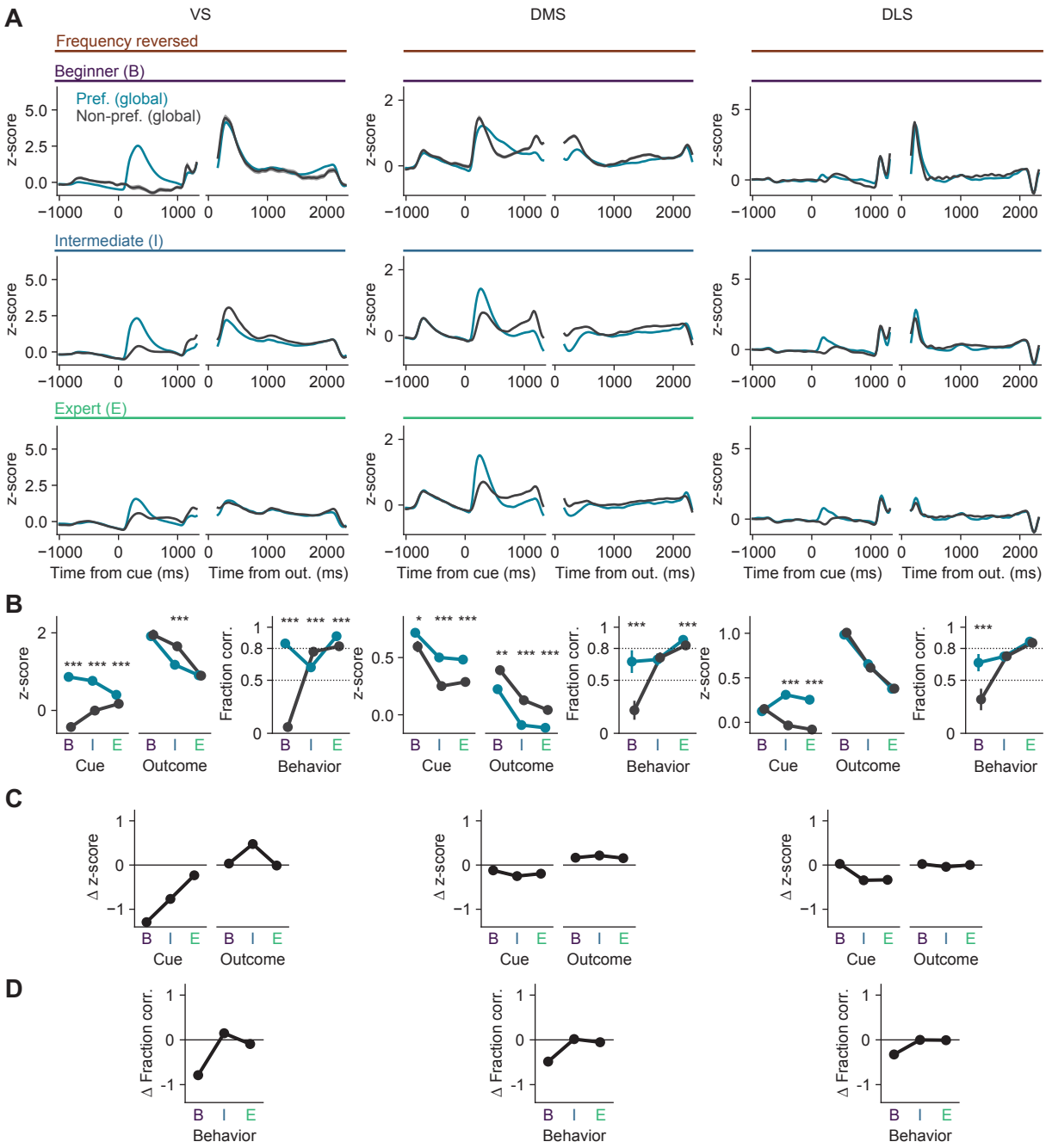


Fig. 21) Dopamine in preferred vs. non-preferred trials by performance levels in the frequency reversed task

A) Average normalized dLight fluorescence in correct trials split by preferred vs. non-preferred trials (global side preference during the whole task rule) for performance levels in the frequency reversed task. Lines show mean across trials, shadings show standard error of the mean across trials (VS, n = 6 animals; DMS, n = 5 animals; DLS, n = 4 animals; see *Appendix* Table A1 in section 6 for number of trials per group). **B)** Average normalized dLight fluorescence, split as in A), averaged across trial epochs (cue, outcome); mean ± standard error of the mean across trials, Wilcoxon rank-sum tests across pooled trials with Bonferroni correction for multiple comparisons. Right-most panel: Average fraction of correct trials, split as in left panel; Chi-squared tests. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. See *Appendix* Table A1 in section 6 for number of trials per group and individual statistical test results. **C)** Difference between average normalized dLight fluorescence in preferred (global) vs. non-preferred (global) trials (Δ z-score), split as in A) and B). **D)** Difference between average fraction of correct trials in preferred (global) vs. non-preferred trials (global) trials (Δ Fraction correct), split as in A) and B).

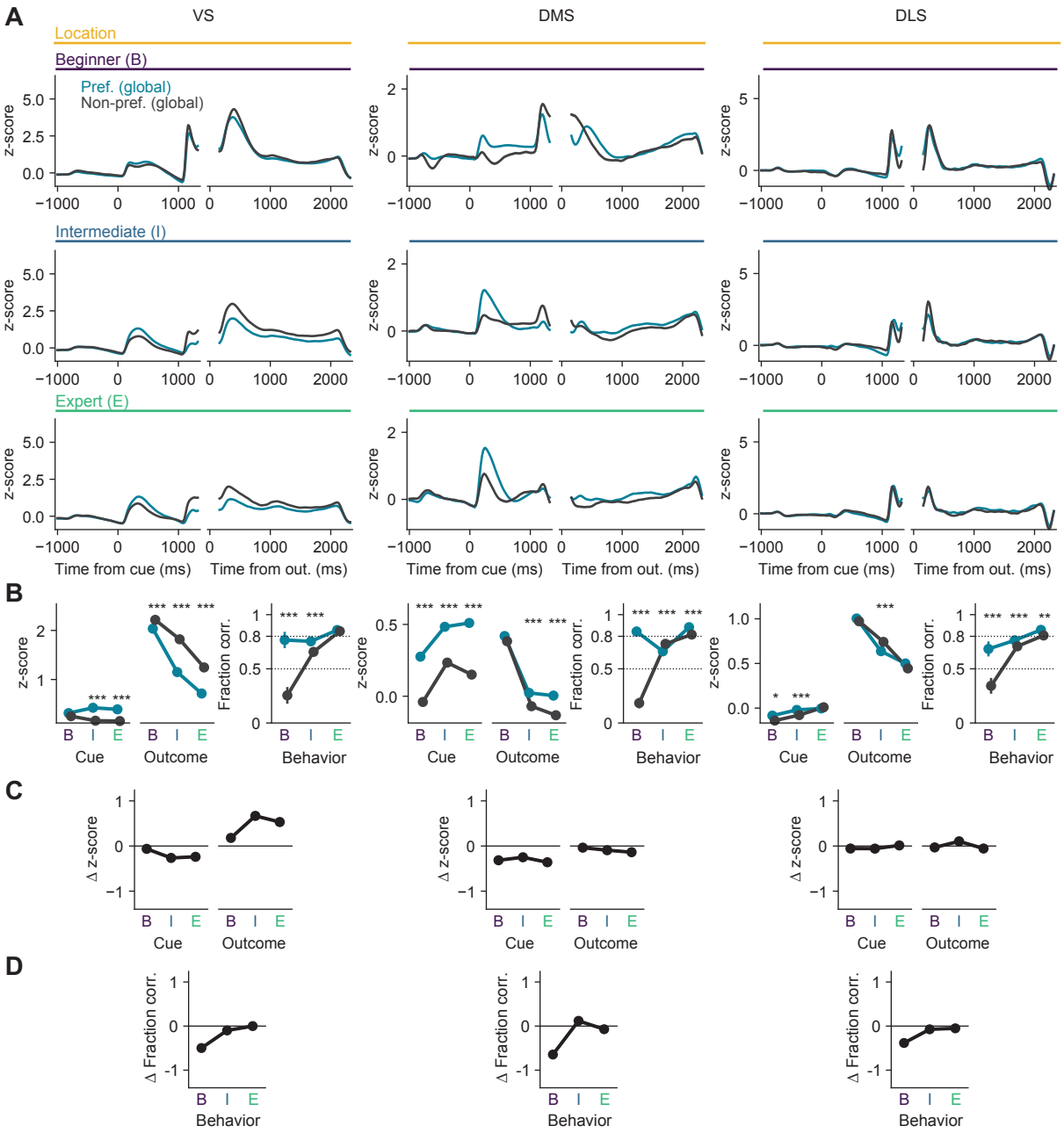


Fig. 22) Dopamine in preferred vs. non-preferred trials by performance levels in the location task

A) Average normalized dLight fluorescence in correct trials split by preferred vs. non-preferred trials (global side preference during the whole task rule) for performance levels in the location task. Lines show mean across trials, shadings show standard error of the mean across trials (VS, n = 6 animals; DMS, n = 5 animals; DLS, n = 4 animals; see *Appendix* Table A1 in section 6 for number of trials per group). **B)** Average normalized dLight fluorescence, split as in A), averaged across trial epochs (cue, outcome); mean ± standard error of the mean across trials, Wilcoxon rank-sum tests across pooled trials with Bonferroni correction for multiple comparisons. Right-most panel: Average fraction of correct trials, split as in left panel; Chi-squared tests. *** p < 0.001, ** p < 0.01, * p < 0.05. See *Appendix* Table A1 in section 6 for number of trials per group and individual statistical test results. **C)** Difference between average normalized dLight fluorescence in preferred (global) vs. non-preferred (global) trials (Δ z-score), split as in A) and B). **D)** Difference between average fraction of correct trials in preferred (global) vs. non-preferred trials (global) trials (Δ Fraction correct), split as in A) and B).

Taken together, dopamine transients were modulated by global reward expectations due to behavioral preference for certain conditions. The preference was experimentally induced after the switch from location to frequency, where trials were split into stay and switch conditions, while the preference was self-selected during the location task and frequency reversed task, where the animals exhibited response biases towards one of the two licking spouts. Regardless of the preference definition, overall, across all subregions and task rules the pattern of dopamine cue and reward responses in preferred compared to non-preferred conditions matched the expected pattern of reward prediction errors, with larger dopamine cue responses in preferred compared to non-preferred conditions and larger dopamine reward responses in non-preferred compared to preferred conditions. Strikingly, in the VS (and to some extent also in the DMS) during beginner sessions the reward prediction error signature partially lagged behind behavioral signs of preference, with dopamine reward responses reflecting behavioral preferences only after a few sessions, suggesting that the signature was corrupted in beginner sessions. The reward prediction error signature was intact again during intermediate and expert sessions.

As already mentioned in section 3.2.4, the preference modulation of dopamine cue responses, together with the distinct behavioral changes associated with the two rule switches, resulted in the diverging pattern of overall dopamine cue responses in the sessions following the two different rule switches shown in Figure 18. In the first session after the rule switch from location to frequency, the dopamine cue response in the VS and DMS rapidly increased and then decreased again (Fig. 18A/B). Since in Figure 18 an average of all correct trials is depicted, both stay and switch trials are included. This average of correct trials included mostly stay trials in the beginning of the sessions when the animals strongly perseverated on the previous rule, and included a mix of stay and switch trials later in the session (Fig. 6). Since the dopamine cue response was larger in stay trials compared to switch trials in the VS and DMS, the average of all correct trials was larger in the beginning of the session than later in the session (Fig. 18A/B). In contrast, in the first session after the switch from frequency to frequency reversed, the mix of correct trials first included both preferred and non-preferred trials and later in the session, when animals entered the response bias strategy, contained mainly preferred trials (Fig. 6). Thus, since the dopamine cue response was larger in preferred trials compared to non-preferred trials, the average dopamine cue response of correct trials increased over the course of the first session after the rule switch from frequency to frequency reversed (Fig. 18C/D).

3.2.6 Previous trial outcome offsets dopamine transients

The analysis in the previous section showed that dopamine transients during both the cue and the reward epoch were modulated by global reward expectations depending on preference for certain conditions. In the VS, these modulations changed on a relatively slow timescale of several sessions, partially lagged behind behavioral changes, and persisted when behavioral differences were not present anymore (Figs. 19-22). Similar modulations were observed to a smaller extent in the DMS, while in the DLS there was only weak modulation by global reward expectations due to behavioral preferences (Figs. 19-22). To test whether dopamine transients across the three subregions were not only modulated by learned global expectations on a slow timescale, but also by faster fluctuating changes in reward expectations (Engelhard et al., 2019), the effect of the outcome in the previous trial on the current trial was examined.

Figure 25 shows the dopamine responses in correct trials split by previous trial outcome, that is, correct (rewarded) previous trials compared to false (unrewarded) previous trials. In the VS and DMS, across all performance levels both dopamine cue and reward responses were larger in trials that followed false trials compared to trials that followed correct trials (Fig. 23A/B). In the DLS, the dopamine reward response showed the same pattern, while the dopamine cue response was not modulated by previous outcome (Fig. 23A/B). Notably, the difference between trials following correct trials versus trials following false trials was already present before cue presentation, especially in the VS (Fig. 23A). The previous outcome modulation of dopamine responses throughout the trial was largely independent of behavior (Fig. 23B). There was a small, yet significant difference in the fraction of correct trials between previous-correct and previous-false trials in intermediate and expert sessions, but not in beginner sessions. Yet, the previous outcome modulation of dopamine responses was present in beginner sessions. Note that while false trials tended to accumulate at the end of the session, the effect of increased dopamine transients after false unrewarded trials was similar when only the first half of the session was considered (data not shown).

Overall, dopamine responses were not only modulated by global reward expectations that changed across sessions (preferred versus non-preferred conditions), but also by fast trial-by-trial fluctuations in reward expectations based on previous trial outcome. However, the pattern of modulation of cue and reward responses was different. Global preference for conditions led to increased dopamine cue responses and decreased dopamine reward responses compared to non-preference (Figs. 19C, 20C, 21C, 22C). Previous false trials without rewards led to a general offset in the form of increased dopamine cue and reward responses in the current trial compared to trials following previous correct trials with rewards (Fig. 23C). This fast trial-by-trial modulation by reward expectations has been previously reported (Engelhard et al., 2019) and is in line with the reward prediction error framework, since both rewards and reward-predicting cues are less expected in trials immediately following unrewarded trials compared to trials following rewarded trials and therefore elicit larger reward prediction errors.

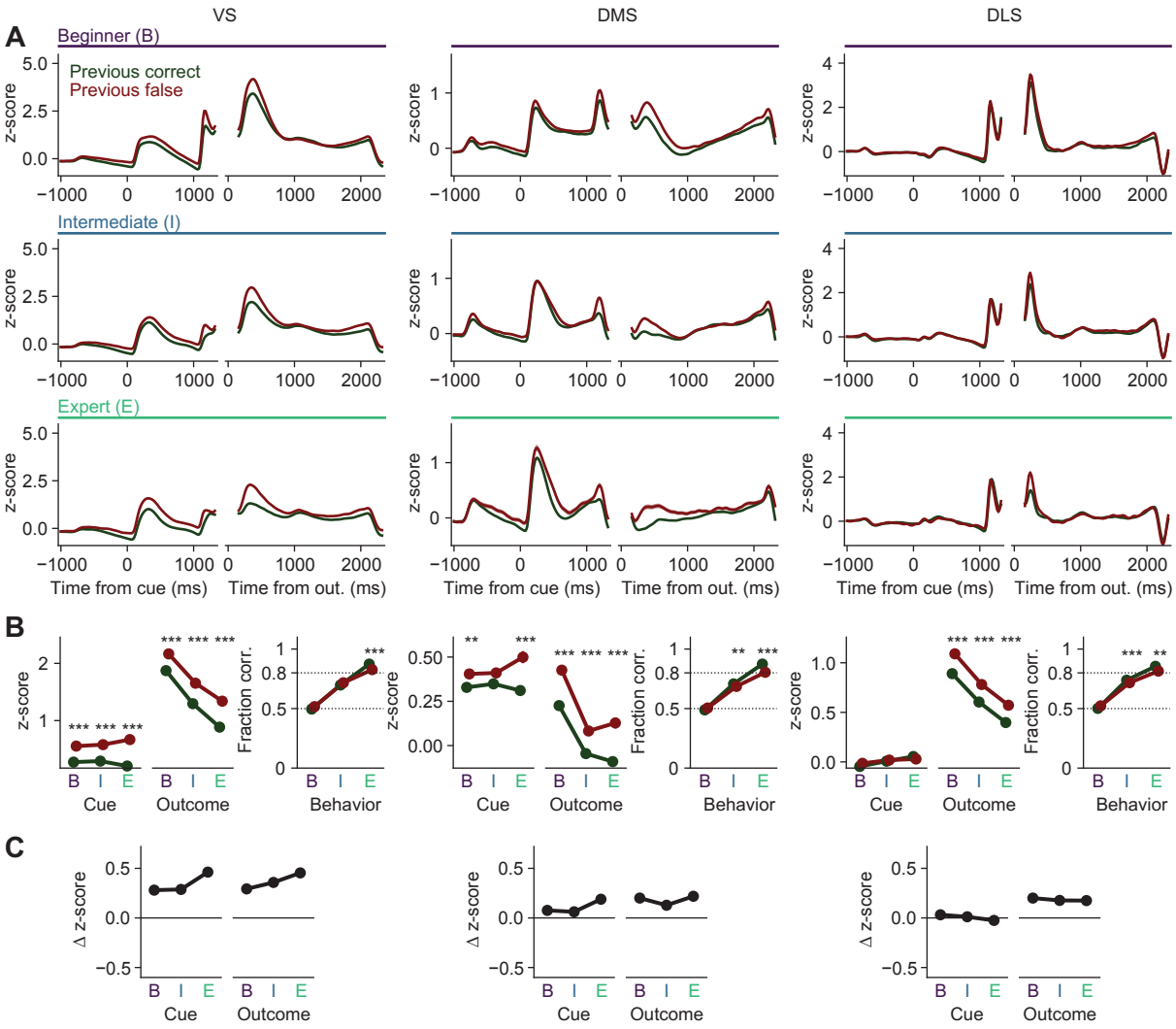


Fig. 23) Dopamine modulation by previous trial outcome
A) Average normalized fluorescence in correct trials split by previous outcome (previous trial correct vs. previous trial false) for performance levels averaged across all task rules. Lines show mean across trials, shadings show standard error of the mean across trials (VS, n = 6 animals; DMS, n = 5 animals; DLS, n = 4 animals; see *Appendix* Table A1 in section 6 for number of trials per group). **B)** Average normalized dLight fluorescence, split as in A), averaged across trial epochs (cue, outcome); mean ± standard error of the mean across trials, Wilcoxon rank-sum tests across pooled trials with Bonferroni correction for multiple comparisons. Right-most panel: Average fraction of correct trials, split as in left panel; Fisher exact tests. *** p < 0.001, ** p < 0.01, * p < 0.05. See *Appendix* Table A1 in section 6 for number of trials per group. **C)** Difference between average normalized dLight fluorescence in preferred vs. non-preferred trials (Δ z-score), split as in A/B), averaged across trials.

3.2.7 Encoding model reveals unique contributions of relevant variables

To test whether the modulation of dopamine signals contained unique contributions of the variables examined so far, when influences of other variables are controlled, a neural encoding model was used. This model allowed to combine behavioral and neural aspects in one analysis to better understand the big picture of dopamine signals in the present task.

A linear regression model (see *Methods*) was used to predict the dopamine signals at every timepoint in the trial (Blanco-Pozo et al., 2021; Musall et al., 2019). Regressors were selected to capture the central effects presented so far. The regressors (Fig. 24A) included the following whole-trial variables: an intercept; current trial outcome; previous trial outcome; the instruction cue weight from the trial-based choice model; multiple-trial-based response preference defined by the response bias weight from the trial-based choice model; multiple-trial-based response preference defined by stay versus switch trials; multiple-session-based global response preference defined by individual response preference during a given task rule. Since the model was fit for every timepoint in the trial (see *Methods*), the intercept represented the overall modulation of the dopamine signal by trial events independent of different trial types (i.e., independent of conditions and outcomes). The current and previous outcome regressors indicated whether or not there was a reward in the current and previous trial, respectively. The instruction cue weight regressor represented the extent to which the animals followed the instruction cue to guide the choice in a given trial (i.e., a measure of task performance, see also *Methods*), as defined by the trial-based choice model. The preference regressors indicated congruence of a certain trial to the preference, that is, whether a trial was currently preferred or not.

Explained variance and unique contributions were obtained from five-fold cross-validation (see *Methods*). The model was fit per session, although a fit per animal yielded similar results for explained variance and unique contributions (data not shown). In all subregions, the full model explained significantly more variance in the dopamine signals compared to a control model, in which all variables were shuffled across trials (Fig. 24B). The shuffled model was equivalent to an intercept-only model representing the average modulation of the dopamine signals by trial events.

To determine unique contributions of regressors, each regressor was shuffled across trials once to quantify the relative reduction in explained variance when a given regressor was eliminated (i.e., the unique contribution of that regressor). All regressors had significant unique contributions averaged across all sessions, except for the local preference regressor in the DMS (Fig. 24C). In all subregions, the contribution of the current trial outcome was qualitatively largest, accounting for the large difference in the dopamine outcome response in correct trials with rewards compared to false trials without rewards (Fig. 14). The contribution of the instruction cue regressor indicates performance-related modulation. Note that the large effect of performance modulation observed in Figure 17 was already accounted for to a large extent by the session-wise fitting. The fact that in the DMS the instruction cue contribution was largest, supports the notion that the DMS reward response showed the fastest adaptation to performance, as seen in Figure 18.

Trial-based preference defined by the choice model had the smallest unique contributions. Note that trial-based preference represented trial history-related biases, which also had a minor influence on choice behavior (Figs. 8-11). Both preference in the form of preference for stay compared

to switch trials as well as global preference in the form of an individual global response preference had significant unique contributions to the average dopamine signals. This suggested a slow emergence of preference-related modulation. Note that the unique contributions in Figure 24 were calculated across all sessions, producing a conservative estimation for the regressors that had their effect mainly during one of the task rules (e.g., preference for stay conditions only in the frequency task). In summary, the encoding model revealed unique contributions of regressors in support of the effects presented so far.

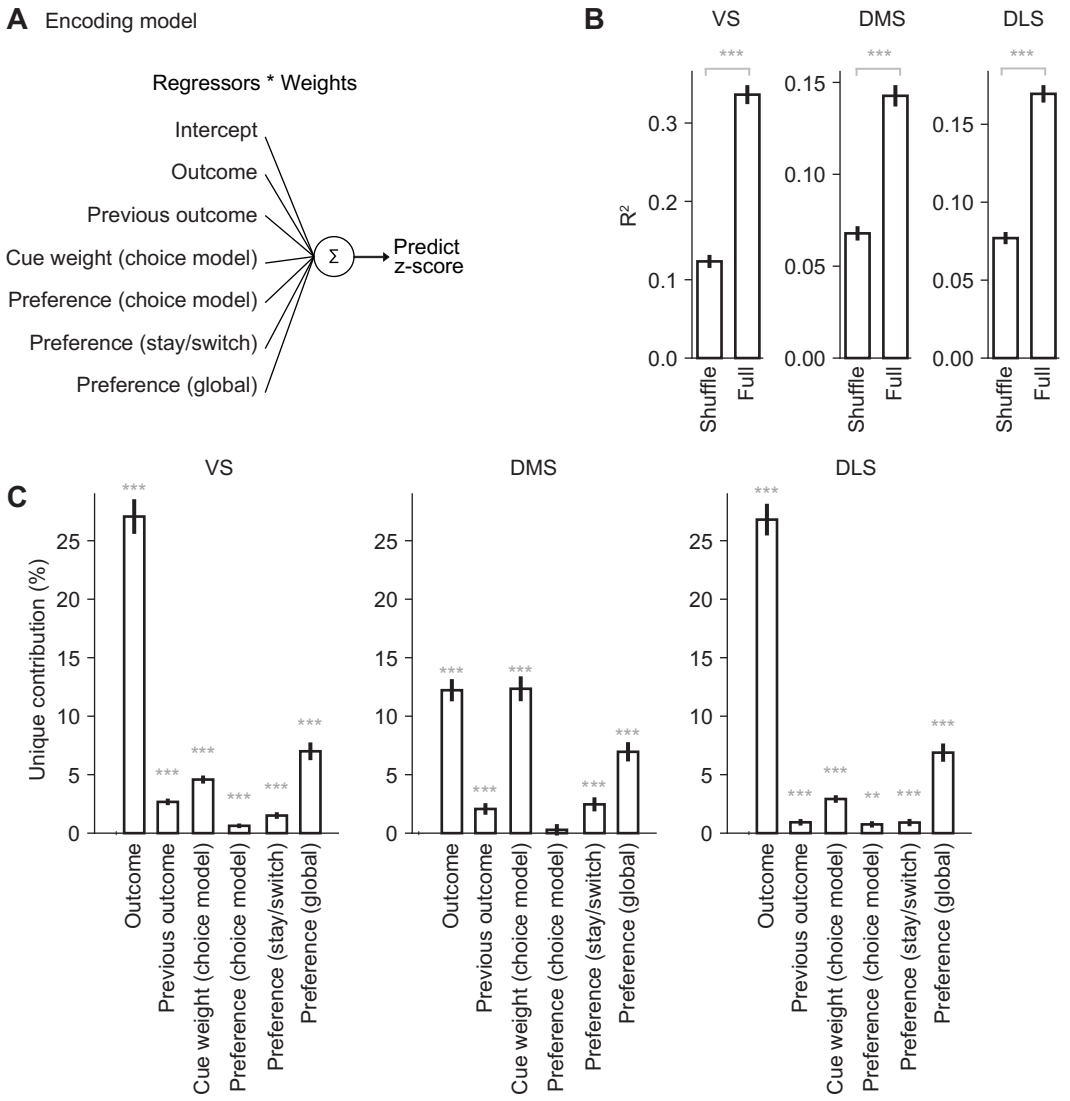


Fig. 24) Encoding model
A) Schematic of the linear regression model fit per session for each timepoint in the trial (see *Methods*). **B)** Cross-validated explained variance (R^2) averaged across the whole trial for all sessions, shown for the full model with regressors shown in A) and for a shuffled model with all regressors shuffled across trials. Mean ± standard error of the mean across sessions, Wilcoxon signed-rank tests across sessions with Bonferroni correction for multiple comparisons. *** p < 0.001, ** p < 0.01, * p < 0.05. **C)** Cross-validated percent unique contribution of each regressor averaged across the whole trial for all sessions. The unique contribution of a given regressor is the relative difference in explained variance between the full model and a model with the regressor shuffled across trials. Mean ± standard error of the mean across sessions, Wilcoxon signed-rank tests for difference from zero across sessions with Bonferroni correction for multiple comparisons. *** p < 0.001, ** p < 0.01, * p < 0.05. See *Appendix* Table A1 in section 6 for number of sessions per group.

The encoding model was fit per session. Fitting per session and per timepoint allowed to examine how the weights were distributed over the trial epochs and how they developed across performance levels and task rules, as described for selected weights in the following. In order to reproduce the effects of preference modulation observed in Figures 19 to 23, Figure 25 shows the regressor weights for the preference regressors (global response preference in the location and frequency reversed task, stay/switch preference in the frequency task; see *Appendix* Fig. A4 in section 6 for all regressor weights). The weight trajectories of these regressors qualitatively supported the previous findings and will first be described for the VS. In the location task, congruency to global response preference modulated the dopamine signals in a positive direction during the cue epoch and in a negative direction during the outcome epoch, but only in intermediate and expert sessions (as seen in Fig. 22). Similarly, in the frequency task, congruency to preference for stay over switch trials modulated the dopamine signal in a positive direction during the cue epoch and in a negative direction during the outcome epoch, with the strongest cue and outcome modulation in intermediate sessions (as seen in Fig. 20). In the frequency reversed task, the dopamine signal was modulated in a positive direction during the cue epoch in beginner and intermediate sessions and in a negative direction during the outcome epoch mostly in intermediate sessions (as seen in Fig. 21). In the DMS and DLS, the dopamine cue and outcome response modulations observed in the weight trajectories also matched the patterns seen in Figures 20 to 22, most clearly visible in dopamine cue response modulations during intermediate and expert sessions in the DMS across all task rules, and dopamine cue response modulations during intermediate and expert sessions in the DLS in the frequency reversed task (as seen in Fig. 21). By and large, the regressor weights therefore confirmed the previously identified pattern of dopamine signatures during cue and outcome epoch in response to preferred versus non-preferred conditions.

Taken together, the neural encoding model summarized results presented in the previous analyses of dopamine session averages. The effects presented in previous sections are based on variables that uniquely modulated the dopamine signals, when other variables were controlled for (Fig. 24). Dopamine signals were uniquely modulated by slowly evolving preferences (global response preference and preference for stay over switch trials) even when the trial-by-trial fluctuating response bias from the trial-based choice model was already accounted for, suggesting that the dopamine signals contained slowly emerging aspects of reward prediction errors that did not change on a trial-by-trial basis. In turn, the unique contribution of trial-by-trial preference (defined by the trial-based choice model) to the dopamine signals was small or even absent (Fig. 24), when slowly evolving preferences were accounted for, suggesting a strong correspondence to the behavioral results, which showed that trial history-related biases overall had a minor influence on choice behavior (Figs. 8-11).

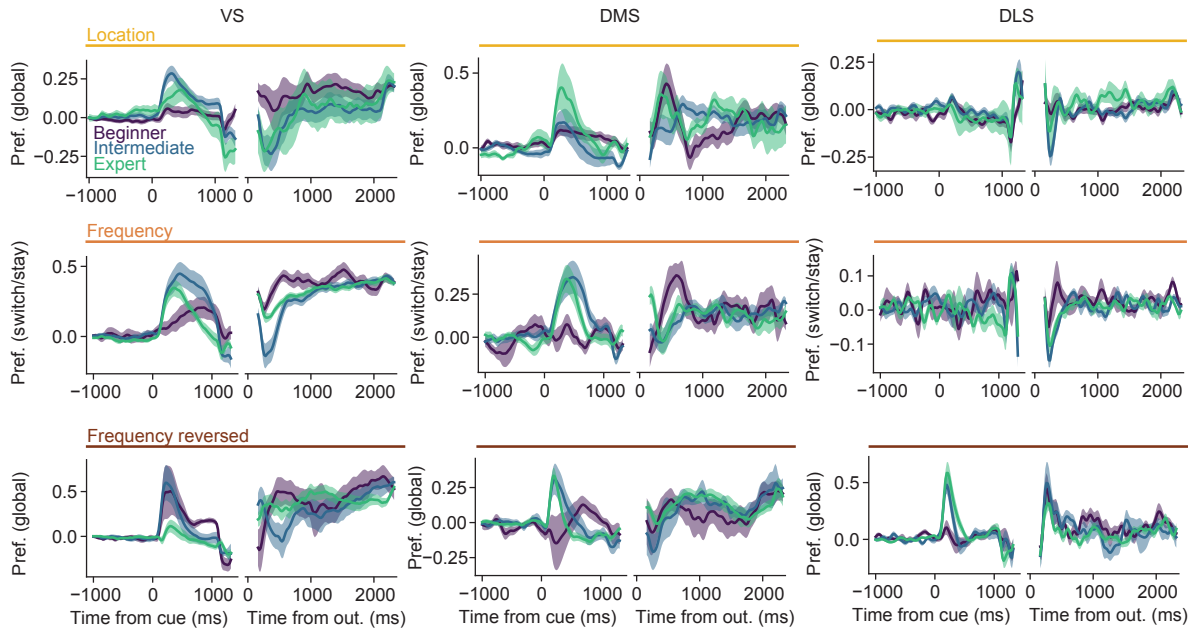


Fig. 25) Encoding model weights (selection)
Normalized regressor weights across the trial for the global preference regressor (location task and frequency reversed task) and the stay/switch preference regressor (frequency task), split by performance levels and task rules. See *Appendix* Fig. A4 in section 6 for weights of all regressors. Lines show mean, shadings show ± standard error of the mean across sessions.

4. Discussion

4.1 Summary of the main results

4.1.1 Behavioral results

In an auditory decision-making task with rule switches (Fig. 3), mice reliably acquired different rule-based associations between auditory instruction cues and instrumental licking responses. Mice started out with strong idiosyncratic response biases and gradually learned to follow the cues as optimal response instructions (Fig. 4). Mice perseverated after rule switches and temporarily kept responding according to the previously learned rule (Fig. 6). Logistic regression models revealed that mice used different strategies to perform the task, depending on the learning state (Figs. 8-11). In particular, in beginner and intermediate-level sessions, but not in expert sessions, mice showed a tendency to repeat previous choices, but no other trial history-related biases (Figs. 8, 10). In expert sessions, mice made cue-instructed decisions on a trial-by-trial basis and their choices were guided by the instruction cue only (Figs. 8, 10). A trial-based choice model with fluctuating weights for response bias and instruction cue parsimoniously explained the animals' choices throughout the learning process (Figs. 10, 11).

4.1.2 Neural results

Dopamine signals were robustly modulated by trial events in the ventral striatum (VS), dorsomedial striatum (DMS), and dorsolateral striatum (DLS). Dopamine signals in all subregions showed signatures of reward prediction error (Schultz et al., 1997) in the form of an activation by rewards, and a deactivation in false trials without rewards (Fig. 14). Dopamine reward responses were inversely scaled to average task performance in all subregions and across all task rules (Fig. 17), in line with previous reports (Hollerman & Schultz, 1998) and in line with the reward prediction error hypothesis, since rewards were better predicted in sessions with higher performance. Dopamine reward responses quickly adapted to changes in performance and resulting reward expectations after rule switches (Fig. 18). In contrast, dopamine responses to the instruction cue were not simply scaled to task performance (Fig. 17), as may have been expected due to previous reports of Pavlovian learning tasks (Fiorillo et al., 2003; Menegas et al., 2017; Tobler et al., 2005) and instrumental learning tasks with varying reward probabilities (Tsutsui-Kimura et al., 2020) or perceptual difficulty (Lak, Okun, et al., 2020). Instead, dopamine cue responses were initially built up even before the animals showed increases in performance (Fig. 16) and were subsequently modulated by behavioral preference for particular conditions (Figs. 19-23).

Dopamine cue and reward responses were modulated by congruency to preference, depending on the current task rule, performance level, and corresponding behavioral strategy. During the frequency task rule, when mice perseverated on the previous rule after the rule switch and preferred conditions that stayed the same over conditions that were switched, VS and DMS showed increased dopamine cue responses and decreased dopamine reward responses in stay conditions compared to switch conditions (Figs. 19, 20). This pattern is in line with the reward prediction error framework, since due to the behavioral preference for stay trials, cues indicating stay trials with high

performance predicted a higher reward probability than cues indicating switch trials with low performance. Likewise, rewards in stay trials were better predicted than rewards in switch trials, and therefore produced smaller reward prediction errors.

Intriguingly, in the VS (and to a smaller degree in the DMS, but not in the DLS) the reward prediction error signature in the dopamine signals partly lagged behind the behavioral signature. The largest difference in performance between stay and switch conditions was directly after the rule switch. Interestingly, the dopamine reward response was equally increased in stay versus switch trials after the rule switch. Even rewards that should have been well predicted, since the contingencies had not changed for stay trials and the animals kept responding to them at high performance levels, elicited a larger dopamine reward response compared to expert sessions before the rule switch. Only over the course of several sessions, the expected reward prediction error pattern emerged, together with an increase in performance. In the VS, during the location task rule and the frequency reversed task rule, a comparable pattern was found for preference according to an individual global response bias (Figs. 21, 22). Importantly, the neural effect lagged behind the behavioral effect, since the largest difference in performance between preferred and non-preferred trials occurred in earlier sessions than the largest difference in dopamine signals.

In addition, dopamine responses throughout the trial were modulated by previous outcome, producing larger responses to both the instruction cue and the reward in trials following errors without rewards compared to trials following correct trials with rewards (Fig. 23). This finding is in line with previous observations (Engelhard et al., 2019) and with the reward prediction error hypothesis, since both rewards and reward-predicting cues are more surprising if they occur in trials following unrewarded error trials.

The main results were summarized in a neural encoding model that revealed significant unique contributions of the relevant predictors and weight trajectories in support of the findings from the session average analyses (Figs. 24, 25).

4.2 Behavioral strategies of mice during learning

4.2.1 No trial history effects except fluctuating response bias

The thorough examination of choice behavior in the present study using a variety of regression models revealed that mouse choices were influenced by previous choices and rewards in beginner and intermediate sessions, but not in expert sessions. In the session-based choice model, several trial history regressors performed equally well, all representing a bias towards repeating previous choices (Fig. 8). A trial-based model with only an intercept and instruction cue regressors performed better than session-based models and better than alternative and more complex trial-based models (Fig. 10), indicating that fluctuations in response bias sufficiently captured trial history effects, while other aspects such as win-stay and/or lose-switch strategies or sensory history played a minor role. The best-performing session-based model included a regressor for value difference, defined as the difference in the reward rate on the left licking spout compared to the right licking spout, exponentially averaged over the previous 10 trials. The value difference regressor therefore represents a model-free reinforcement learning strategy of choosing the spout with the larger reward probability based on recent choices and rewards, independent of the instruction cue. Interestingly, this regressor only had a significant contribution during beginner and intermediate sessions, when mice showed individual global response biases. Since the values of choices were not experimentally manipulated and value difference only reflected the recent history of choices and rewards, no causality could be established, whether the value differences were a consequence of the choices or vice versa. The most parsimonious explanation for the behavior was therefore a trial-by-trial fluctuation in choice bias (as confirmed by the trial-based model). The value difference regressor was not relevant in expert sessions, suggesting that mice did not follow a purely economic strategy of just picking the higher-valued licking spout when they had fully learned the task. Instead, in expert sessions mice followed the cues as response instructions. This was most likely a consequence of the controlled task design without external reward manipulations, since picking the higher-valued spout according to the history of rewards was not an advantageous strategy and only following the instruction cue was the optimal strategy in the present task.

Several previous studies have reported trial history-dependent biases in different tasks and across species (Abrahamyan et al., 2016; Akrami et al., 2018; Busse et al., 2011; Lak, Hueske, et al., 2020; Roy et al., 2021). Trial history effects are amplified by low perceptual confidence (Lak, Hueske, et al., 2020) and are very specific to the behavioral task. For example, Akrami and colleagues (2018) reported sensory history effects in a task that required rats to compare consecutive sensory stimuli. In the present study, only the history of previous choices had an influence on mouse choices during beginner sessions and sessions with intermediate performance, but not expert sessions, likely because other factors were well-controlled. Thus, in the behavioral paradigm presented here, mice made trial-by-trial decisions based on the instruction cue and in the expert state were not influenced by the recent history of choice and rewards.

4.2.2 Behavior in the present task in relation to learning theory

Instrumental behavior is often classified as goal-directed or habitual (see Fig. 1), or along the reinforcement learning distinction of model-free and model-based learning. Relating the mouse behavior in the present study to these concepts may help to compare the results to other studies, even though the interpretations of the results presented here do not depend on a formal classification along these lines.

A hallmark of goal-directed behavior is a sensitivity to devaluation of rewards or contingency degradation, while habitual behavior is insensitive to these manipulations (Balleine & O'Doherty, 2010). Insensitivity to changes in contingencies is observed through perseverative behavior, which indicates that outcomes no longer influence behavior. The distinction between habitual and goal-directed behavior may often be difficult, since there is no formal definition of how long exactly an animal needs to show perseveration after changes in contingencies for a behavior to be classified as habitual. Nevertheless, the relatively fast strategy switches observed after rule switches in the present behavioral task could be regarded as evidence for goal-directed behavior. After the rule switch from frequency to frequency reversed, when all contingencies had been reversed, the animals stopped perseverating after around one third of the session (Fig. 6), which indicates that they were sensitive to changes in the outcome.

Goal-directed behavior is often equated with model-based learning (Drummond & Niv, 2020). In this study, mice showed a tendency to repeat previous choices, that is, a trial history bias, during beginner and intermediate sessions, but not in expert sessions (Figs. 8, 10). The presence of trial history biases during beginner and intermediate-level sessions could be regarded as evidence for a model-free strategy, since the values of the two licking spouts (and corresponding choices) were updated on a trial-by-trial basis according to the rewards received. Likewise, the absence of trial history effects during expert sessions could indicate a model-based strategy, since mice did not rely on trial-by-trial changes of choice values in expert sessions. This would mean that the mice progressed from a suboptimal model-free strategy to an optimal model-based strategy throughout the learning process. In expert sessions, mice had learned a strong association and made trial-by-trial decisions only on the basis of the instruction cue only.

4.3 Reward prediction errors during acquisition of instrumental associations

4.3.1 Reward response inversely scaled to performance, cue response not

A central question of this study was, how dopamine cue responses and dopamine reward responses evolved in correlation with task performance during acquisition of novel instrumental associations. The present behavioral paradigm was different compared to previous studies that used Pavlovian tasks (Menegas et al., 2017) or instrumental tasks with varying reward probabilities (Tsutsui-Kimura et al., 2020) or perceptual difficulty (Lak, Okun, et al., 2020). In the present task, the receipt of a reward depended on a correct decision and corresponding action, which was determined by the current task rule. Thus, the receipt of a reward depended on the current learning state and the strength of the association between instruction cue and required action to achieve the desired outcome, but not on external reward probabilities. The present paradigm therefore enabled to examine how dopamine signatures evolved during the previously understudied acquisition of associations in a decision-making task in the absence of external reward manipulations.

Over the course of learning, the dopamine response to the reward in all subregions inversely scaled with task performance, as measured by the fraction of correct trials (Fig. 17), a phenomenon reported previously in dopamine neurons in the ventral tegmental area of monkeys during a choice task (Hollerman & Schultz, 1998). With increasing performance rewards became better predicted and the dopamine reward response decreased. This effect is in line with the framework of reward prediction errors (Schultz et al., 1997) and with numerous observations that dopamine reward prediction errors inversely scale with reward expectations (Glimcher, 2011; Watabe-Uchida et al., 2017). Thus, this finding reproduced previous evidence stemming mainly from Pavlovian and simple instrumental tasks during the acquisition of the present decision-making task. With fast increases after rule switches (Fig. 17), dopamine reward responses rapidly adapted to changes in task performance and reward expectation, which indicated that the animals had a firm knowledge of the task and strong expectations of the average number of rewards that they received in the task. This is in accordance with findings from the behavioral analysis, which revealed that the mouse behavior in the present task was well-controlled with minor influences of trial history-related biases (as discussed in section 4.2.1).

As described in the introduction (section 2.4.2), compared to Pavlovian tasks, the dopamine response to the instruction cue and its interpretation in the context of the reward prediction error framework inherently becomes more complicated in instrumental tasks, in which the receipt of a reward is not fully (or probabilistically) predicted by the cue, but is conditional on a correct choice. In Pavlovian tasks, the dopamine cue response simply scales with reward expectation (Fiorillo et al., 2003; Tobler et al., 2005). In instrumental tasks with varying external reward probabilities, the dopamine cue response scales with reward expectations when animals are fully trained. In particular, when animals choose between cues with different reward probabilities and thus different values, dopamine neurons have been shown to encode chosen value during the cue epoch (Lak et al., 2016; Morris et al., 2006). Based on these previous findings and the reward prediction error theory, one hypothesis was that in the present task without variation in external reward probabilities, the dopamine cue response should scale with task performance, analogous to the inverse scaling observed in the dopamine reward response. Cue values should scale with performance, since the external reward

probabilities were constant and only the state of the learned association (as indicated by the performance) determined the average reward probability. With increasing performance, the auditory cues should become better predictors of rewards and the dopamine cue responses should increase. Contrary to this hypothesis, in the data presented here the dopamine cue response did not simply scale with task performance analogous to the scaling in the dopamine reward response (Fig. 17). During the initial task rule, there was a slight increase in the dopamine cue response with increasing performance, most clearly observed in the DMS (Fig. 17). However, in stark contrast to the dopamine reward response, the dopamine cue response was not reset after rule switches, when performance dropped to beginner level. Thus, the dopamine cue response was on average not scaled to global reward expectation across learning. Further analyses revealed that the dopamine cue response was scaled to reward expectation depending on behavioral preference for particular conditions (Figs. 19-22, discussed below).

Overall, the data presented here suggest that during the acquisition of instrumental associations, when external reward probabilities are constant and the receipt of rewards depends solely on the current learning state, the average dopamine reward response faithfully tracks global reward expectation. In contrast, the dopamine cue response on average does not reflect global reward expectation analogous to the dopamine reward response, but instead depends on behavioral preference for particular conditions (see below).

4.3.2 Partial lag of dopamine signature relative to behavior

Another central question of this study was, how the predictive properties and the reinforcing properties of dopamine reward prediction error signatures contribute to the acquisition of instrumental associations (Coddington & Dudman, 2018; Pan et al., 2021). In the present task, dopamine transients were modulated by global reward expectations due to behavioral preference for particular conditions (Figs. 19-22). The behavioral preference was experimentally induced after the switch from location to frequency, when animals preferred stay over switch conditions, while the preference was self-selected during the location task and frequency reversed task, when animals had individual response biases towards one of the two licking spouts. Overall, across all subregions and task rules the pattern of dopamine cue and reward responses in preferred compared to non-preferred conditions matched the expected pattern of reward prediction errors, with larger dopamine cue responses in preferred compared to non-preferred conditions and larger dopamine reward responses in non-preferred compared to preferred conditions. Notably, the neural encoding model revealed that dopamine signals were uniquely modulated by slowly evolving preferences (global response preference and preference for stay over switch trials) even when the trial-by-trial fluctuating response bias from the trial-based choice model was already accounted for, while in turn the unique contribution of trial-by-trial preference was small or even absent when slowly evolving preferences were accounted for (Fig. 24). This finding emphasizes the role of slowly emerging dopamine reward prediction error signatures and corresponds well to the behavioral results, which showed that trial history-related aspects in the form of trial-by-trial fluctuations of response biases overall had a minor influence on choice behavior (Figs. 8-11, discussed in section 4.2.1).

Strikingly, in the VS during beginner sessions the reward prediction error signature partially lagged behind behavioral signs of preference, with dopamine reward responses reflecting behavioral preferences only after a few sessions, suggesting that the signature was corrupted in beginner sessions. The reward prediction error signature was intact again during intermediate and expert sessions. What may seem as a deviation from the reward prediction error framework at a first glance, may also be interpreted as a mechanism of learning, when the predictive properties and reinforcing properties of dopamine reward prediction errors are regarded separately.

The fact that after rule switches the dopamine reward response for preferred and non-preferred conditions was of the same amplitude, while the dopamine cue response was already modulated, suggests that at this point there was no association between the correct action that triggered the reward and the evaluation of the outcome. If there had been an association, an unexpectedly correct non-preferred trial should have led to a larger dopamine reward response than an expectedly correct preferred trial, especially since the increased dopamine cue response suggested that there was a reward prediction that discriminated between preferred and non-preferred trials. Thus, it can be inferred that during these beginner sessions the prediction was intact, but the reinforcement was not. In general, the dopamine response to the reward quickly and accurately tracked the average reward rate, both on a long timescale (dopamine reward responses increased after rule switches) and on a short timescale (modulation by previous outcome on a trial-by-trial basis), but during beginner sessions after the rule switch the dopamine reward response did not track the association to the cue. Since in the present task cue and reward values did not depend on external reward probabilities, but on the learned association (i.e., the state of learning), the timescale of the emergence of the full reward prediction error signature matched the timescale of the behavioral learning curves. This phenomenon can be viewed from two angles. The reward prediction error can be regarded purely as a consequence of the behavior, which questions the role of these signals in learning (Coddington & Dudman, 2018). Alternatively, the reward prediction error signature can be regarded as corrupted in early sessions after the rule switch, when mice perform at beginner level (i.e., when the learned association is also corrupted), and it can be regarded as intact again during intermediate-level sessions, when learning takes place (i.e., when the learned association becomes intact again). The larger dopamine reward responses in unexpectedly correct non-preferred trials relative to expectedly correct preferred trials could act as reinforcing and be a cause for, rather than a consequence of learning.

Previous studies have reported lags between behavioral signs of learning and dopamine reward prediction error signatures in Pavlovian learning (Coddington & Dudman, 2018; Lak et al., 2016; Menegas et al., 2017) and instrumental learning (Lak et al., 2016). This has led authors to question the causal role of these dopamine signatures in learning (Coddington & Dudman, 2018). In Pavlovian tasks, most studies found a delayed increase in the dopamine cue response relative to increases in approach behavior. In the data presented here, during instrumental learning the dopamine cue response (i.e., predictive aspect of the reward prediction error signature) adapted quickly to behavioral preferences after rule switches, but the reward prediction error pattern in the dopamine reward response (i.e., reinforcing aspect) lagged behind. Since during low-performance

beginner sessions after rule switches the reinforcing aspect was corrupted, while the predictive aspect remained intact (i.e., quickly adjusted to behavioral preferences), the present data show a strong correlation of the reinforcing aspect with learning, while the predictive aspect was largely independent of learning. This interpretation is in accordance with previous optogenetic studies that found a causal role of reinforcement, but not prediction, in Pavlovian learning (Pan et al., 2021) and during learning as updating of probabilistic reward values (Lak, Okun, et al., 2020). In the present study, no causality can be established. Ultimately, more experiments including manipulations are required to test whether the predictive and reinforcing aspects of reward prediction error signatures have a causal impact on the acquisition of instrumental associations (e.g., using optogenetic manipulations, see also section 4.5).

4.3.3 Other reward prediction error signatures

Dopamine signals in all striatal subregions showed signatures of reward prediction error (Schultz et al., 1997), observed in form of a central reward prediction error hallmark, namely, an activation by rewards, and a deactivation in unrewarded trials (Fig. 14). The present study revealed several other dopamine signatures that are in line with the reward prediction error framework. Alongside signatures discussed above (sections 4.3.1 and 4.3.2), another aspect was found in the dopamine spout response. The dopamine response during the spout epoch, that is, when licking spouts were moved into the animals' reach after the cue offset, evolved across pre-training and the first sessions in the task in a way that can be explained by reward prediction errors (Fig. 16). During pre-training there were no instruction cues, only one of the two licking spouts was presented to the animals, and a reward was triggered by an instrumental lick on the presented spout. The animals therefore obtained a reward in close to 100 percent of the trials. The spout response increased across pre-training sessions, when rewards were almost certain, and then slowly decreased over the first sessions in the task, when animals performed at beginner level of around 50 percent correct trials and thus obtained rewards in approximately half of the trials. Therefore, the spout movements were certain predictors of the reward during pre-training, while they were not predictive of rewards during the task. The increase in dopamine spout responses during pre-training and the decrease during early task sessions are therefore in line with the reward prediction error framework (Schultz et al., 1997) and suggest that the spout movements acted as a reward-predictive Pavlovian cue during pre-training. The dopamine response to the spouts faded, once they were no longer predictive of rewards. This phenomenon was observed across all three striatal subregions.

4.4 Commonalities and differences between striatal subregions

4.4.1 Common reward prediction error coding

Multiple aspects of reward prediction error coding in dopamine signals were common to all three striatal subregions in this study. First, dopamine transients in all subregions exhibited positive prediction errors in response to rewards and negative prediction errors in unrewarded trials (Fig. 14, *Appendix* Fig. A1). Second, all subregions exhibited an inverse scaling of the dopamine reward response to average task performance and therefore average reward expectation (Fig. 18, discussed in section 4.3.1). Third, all subregions showed a fading of dopamine spout response when spout movements were no longer predictive of rewards (Fig. 16, discussed in 4.3.3). Finally, while dopamine responses were not modulated equally strongly by behavioral preference for particular conditions across the striatal subregions, all observed preference modulations were in accordance with reward prediction errors (Figs. 19-23, discussed in 4.3.2).

In the data reported here, in false trials without rewards all subregions exhibited negative prediction errors in the form of fluorescence levels below baseline (Fig. 14, *Appendix* Fig. A1). A previous study comparing the same striatal subregions as the present study in mice performing an olfactory decision-making task did not find negative prediction errors and an overall positively shifted pattern of reward prediction errors in the DLS (Tsutsui-Kimura et al., 2020). While in the present study the DLS exhibited negative prediction errors, the second notion of a positively shifted pattern of prediction errors was not formally tested. One technical difference that may explain the diverging finding is that the aforementioned study recorded axonal activity using a fluorescent calcium sensor as a proxy of neuronal activity in dopaminergic axons, while in this study dopamine fluctuations were measured directly using a fluorescent dopamine sensor. Different kinetics of the sensor and differences in the behavioral tasks may both contribute to the diverging findings.

Interestingly, negative reward prediction errors were not inversely scaled to task performance (*Appendix* Fig. A1). This is in contrast to previous reports of symmetric dopamine reward prediction errors, which were inversely scaled to reward expectations both during rewards and during negative outcomes without rewards (Hart et al., 2014). One reason for this diverging result may be the instrumental behavior in the present study, where rewards are dependent on correct actions. When animals are experts, they may be more aware of their errors even before the outcome is revealed and therefore have lower reward expectations and equally negative prediction errors across performance levels.

4.4.2 Differences in reward prediction error signatures

Despite strong similarities in reward prediction error coding across striatal subregions observed in this study, there were differences between subregions in the temporal dynamics and reward prediction errors signatures across learning.

The dopamine signals in the VS differed from signals in the other subregions mainly by showing the clearest effect of a corrupted reward prediction error signature after rule switches (Figs. 19-22). The reinforcing aspect of the reward prediction error signature during the reward lagged

behind the behavioral signature and therefore was corrupted during beginner sessions after rule switches, when the animals' task performance was low. The signature correlated with learning, since it was intact again in intermediate sessions, when the animals were improving in performance. These aspects support the notion of dopamine signals in the VS being involved in value learning and are in line with the role of the VS as a critic that learns from reward prediction errors in actor/critic reinforcement learning models (O'Doherty et al., 2004).

In several aspects, the dopamine signals in the DMS were different from those in the other subregions. First, dopamine in the DMS showed the fastest adaptations of the dopamine reward response to reward expectations (Fig. 18). Second, only in the DMS the dopamine cue response was larger than the dopamine reward response across all performance levels and task rules (except during beginner sessions in the location task; Fig 17). Third, in the DMS the dopamine reward response was generally small and in expert sessions the dopamine reward response was vanished (in the location task), or even decreased below zero (in the frequency and frequency reversed task). The final aspect has been described before in studies that did not find strong dopamine reward responses in the DMS during instrumental behavior in the fully trained state (Blanco-Pozo et al., 2021; Brown et al., 2011; Parker et al., 2016; Tsutsui-Kimura et al., 2020). By examining the whole learning process and not just the fully trained state, the present study showed that dopamine reward responses can be observed in the DMS in training states with low behavioral performance. A previous voltammetry study reported that dopamine in the DMS did not show strong reward responses, except after changes in reward contingencies (Brown et al., 2011). This is in line with the strong and transient modulation of dopamine reward responses in the DMS after rule switches in the present study (Fig. 18). Taken together, dopamine signals in the DMS could be used to facilitate flexible behavior after rule switches in accordance with the assumed role of the DMS in goal-directed behavior and behavioral flexibility (Grospe et al., 2018; Redgrave et al., 2010).

Finally, the signals in the DLS were different from the other subregions. Dopamine signals in the DLS did not show strong cue responses (Fig. 17), did not show strong modulation by global preference to particular conditions (Figs. 19-22), showed previous outcome modulation only in the reward epoch (Fig. 23), and maintained the largest relative spout responses compared to the other subregions throughout the experiment (Fig. 17). These aspects may be correlates of a distinct role of dopamine signals in the DLS being involved in directly reinforcing actions through policy learning (Tsutsui-Kimura et al., 2020), in accordance with the proposed role of the DLS in skill learning and habit formation (Redgrave et al., 2010) and in line with the idea of the DLS as an actor that learns from performance errors in actor/critic reinforcement learning models (O'Doherty et al., 2004).

4.4.3 Different dynamics of dopamine fluctuations

Apart from differences in dopamine reward prediction errors across the three striatal subregions, dopamine fluctuations in the three subregions also showed distinct temporal dynamics. Dopamine peaks were broadest in the VS, narrower in the DMS, and narrowest in the DLS (Fig. 14). Wei and colleagues (2021) recently reported similar systematic variation in dopamine fluctuations across the same three subregions and showed that the temporal dopamine dynamics were related to

corresponding differences in reward integration and discounting, supporting different roles in decision-making. The subregion-specific aspects of dopamine signals presented in this study are largely in line with this proposal. The partial lag of dopamine reward prediction errors relative to preference behavior in the VS (Figs. 19-22) could be a consequence of longer reward integration. The relatively fast adaptations of DMS dopamine responses to rewards after rule switches (Fig. 18) could be a consequence of faster reward integration. The weak modulation by preference for particular conditions and the absence of lags between preference behavior and dopamine reward prediction error signatures in the DLS (Figs. 19-22) could be due to the fastest reward integration in this subregion.

4.5 Future directions

Further experiments are required to elucidate the causal mechanisms behind dopamine signatures described in this study. Temporally precise manipulation of specific dopamine projections using optogenetics could establish causal relationships between reward prediction error signatures and learning. If the reward prediction error signature of preferred versus non-preferred conditions causes learning and is corrupted after rule switches, exogenously restoring it with optogenetic stimulation in early sessions after rule switches, when the signature seems to be corrupted, should enhance learning. This concrete hypothesis could be tested in future experiments. Pan and colleagues (2021) recently showed that stimulation of dopamine neurons in mice was sufficient to induce Pavlovian approach behavior when it replaced rewards, but not when it replaced reward-predicting cues, suggesting that the stimulation was sufficient for reinforcement, but insufficient for prediction. Another study found a similar causal effect on learning behavior by optogenetically manipulating the reinforcing aspect of reward prediction errors in dopamine neurons during the reward, but not by manipulating the predictive aspect during the cue, in an instrumental task where learning was defined as updating of probabilistic reward values (Lak, Okun, et al., 2020). In the present study, in the VS the reinforcing aspect of dopamine reward prediction errors during the reward was correlated with learning, while the predictive aspect during the cue was independent of learning. Similar to the aforementioned studies, the causal nature of this effect could be probed with optogenetics in future experiments.

In order to further the understanding of the many functions of dopamine, the behavioral task and dopamine measurement techniques presented here could be used to record dopamine signals not only in the striatum, but also in cortical areas such as the medial prefrontal cortex and the orbitofrontal cortex. Since in rodents cortical dopaminergic projections have a much lower density than striatal projections (Poulin et al., 2018), the fluorescent sensor used in this study may not be sufficiently sensitive to record presumably small dopamine signal amplitudes in the cortex, but more sensitive sensors may be broadly available soon (Patriarchi et al., 2020). Little is known so far about the time course of dopamine signals in the cortex during decision-making. The medial prefrontal cortex is assumed to be a central hub for executive control and not only receives dopaminergic inputs, but also sends feedback connections to dopamine neurons in the midbrain (Carr & Sesack, 2000) as well as to the DMS and VS (Hunnicutt et al., 2016). It is therefore ideally positioned to play a central role in the present behavioral paradigm. Likewise, the orbitofrontal cortex may be critical for the behavior in the present study. The orbitofrontal cortex not only is a dopaminergic projection target (Menegas et al., 2015), but also sends projections to midbrain dopamine neurons and to the striatum (Hunnicutt et al., 2016).The orbitofrontal cortex encodes expected outcome values and is required for flexible behavior in response to contingency changes (Schoenbaum et al., 2009), possibly because it conveys value signals, which are necessary to compute reward prediction errors, to midbrain dopamine neurons (Takahashi et al., 2011). Thus, investigating dopamine signals in prefrontal cortex and orbitofrontal cortex may help to further the understanding of the mechanisms of decision-making and the acquisition of instrumental associations.

5. Methods

5.1 Animal procedures

5.1.1 Animals

All animal procedures were authorized by the local government (Regierung von Oberbayern, license number Az. 55.2-1-54-2532-119-2017). Animal health was examined and scored every day. Wild-type male mice (C57BL/6J, Charles River) were used for all experiments. At the time of surgery, mice were 8-10 weeks old. Animals in the experiments were housed in single cages on a reverse 12-hour light/dark cycle (i.e., dark during the day) and had ad libitum access to food. Mice had ad libitum access to water, except during behavioral experiments (see section 5.2.2).

5.1.2 Virus injection and fiber implantation

Mice were initially anesthetized with 2 percent isoflurane and transferred to a stereotaxic frame (Neurostar). Analgesia (Novaminsulfon 200 mg/kg body weight) was injected subcutaneously prior to the surgery. Throughout the surgery, anesthesia was maintained at 0.8-1.5 percent isoflurane. Body temperature was controlled with a thermometer and adjustable heating pad and respiration was visually monitored. Hair above the skull was removed with a shaver and the skin above the skull was disinfected with 70 percent ethanol. Local anesthetics (2 percent Lidocain solution) were injected subcutaneously and the skin above the skull was excised. The skull was cleaned thoroughly with 0.9 percent sodium chloride solution and roughened with forceps in preparation for implantation. Guided by automated navigation software (Neurostar) for correction of tilt and scaling of the skull, a 0.6-millimeter craniotomy was performed above the target location of virus injection and fiber implantation.

For expression of the fluorescent dopamine sensor dLight (Patriarchi et al., 2018), 200 nanoliters of pAAV5.hSyn.dLight1.2 (4×10^{12} particles/milliliter) were injected unilaterally with a glass-capillary nanoinjector (Neurostar NanoW). pAAV-hSyn-dLight1.2 was a gift from Lin Tian (Addgene viral prep #111068-AAV5; http://n2t.net/addgene:111068; RRID: Addgene_111068). The virus was injected in one of the following target regions: lateral ventral striatum (VS; Bregma + 1.3 millimeters anterior, +/- 1.8 millimeters lateral, + 4.3 millimeters ventral), dorsomedial striatum (DMS, Bregma + 0.7 millimeters anterior, +/- 1.3 millimeters lateral, + 2.6 millimeters ventral), or dorsolateral striatum (DLS; Bregma + 0.5 millimeters anterior, +/- 2.5 millimeters lateral, + 2.8 millimeters ventral). 7 mice were implanted in the VS, 5 mice were implanted in the DMS, and 4 mice were implanted in the DLS. One animal with VS implantation was excluded from the neural analysis due to low dLight signal amplitude. Two animals with VS implantation showed loosened implants half-way through the experiment and were included up to the point that dLight signals became unstable. Left and right hemispheres were counterbalanced across animals for each target region.

The glass capillary with a tip diameter smaller than 50 micrometers was lowered and retracted at 0.5 millimeters per minute. The virus was injected at a rate of 50 nanoliters per minute. The glass capillary was retracted after 10 minutes of diffusion time. Ready-to-implant 1.25-millimeter optic fiber ferrules (Thorlabs CFMXD05 or CFMXD04) or equivalent custom-built ferrules (using Thorlabs FP400URT, CFX440, NOA63) with a fiber diameter of 400 micrometers and numerical aperture of 0.5

were implanted 200 micrometers above the injection site. The length of the implanted fibers was approximately 0.5 to 1 millimeter longer than the dorsoventral implantation coordinate. To minimize tissue trauma, the fiber was iteratively lowered 200 micrometers and retracted 100 micrometers at a speed of 2 millimeters per minute until the implantation site was reached. The ferrule was fixed to the skull with light-curing adhesive (OptiBond All-in-one) and dental cement (Tetric EvoFlow). A custom-made metal bar for head fixation was fixed to the skull posterior to the fiber implant. The implant was covered with black nail-polish to reduce ambient light contamination during imaging. The ferrule was covered with a plastic cap. After the surgery animals rested in their home cage on a heating pad for 30 to 60 minutes. Animals received further analgesia via subcutaneous injections of long-acting meloxicam (1.5 milligrams per kilogram body weight) immediately after the surgery and on three days following the surgery.

5.1.3 Histology and anatomy

After behavioral and imaging experiments (see sections 5.2 and 5.3), animals were perfused with 4-percent paraformaldehyde. Brains were post-fixated for 24 hours with fiber implants still in place to guarantee visibility of the fiber tract in histological slices. Brains were sliced in coronal sections at 120 micrometers using a vibratome and covered with mounting medium (VectaShield) containing DAPI for visualization of cell nuclei. Histological slices were imaged using a confocal Microscope (Leica SP8) with 10-fold magnification. Confocal images were overlayed with sections from the mouse brain atlas (Paxinos & Franklin, 2001) for verification of implantation locations (see Fig. 12).

5.2 Behavioral procedures

5.2.1 Behavioral setup

Behavioral experiments were performed in sound-attenuated boxes (Med Associates). Using the implanted custom head bar, mice were head-fixed and rested in a body tube, where they were able to move their limbs. Two drinking spouts could be moved into and out of reach of the animals' tongue using servo motors (Turnigy TGY 313C) and a micro-controller (Arduino Uno Rev3). When positioned within the reach of the tongue, the spouts were approximately 4-8 millimeters apart and a few millimeters anterolateral to the nose. The spouts were used both as response devices and for water reward delivery. Spout contacts of the tongue (i.e., lick responses) were monitored as threshold crossings of the metal-to-water junction potential (Hayar et al., 2006). Water rewards were dispensed from the spouts using a TTL-controlled syringe pump (New Era Pump Systems NE-500). Speaker drivers and electrostatic speakers for ultrasonic sound production (Tucker-Davis Technologies ED1 and ES1) were used for cue presentation. The speakers were positioned 10 centimeters to the left and to the right of the animals' ears at an angle of 15 degrees towards the front. A screen with a diagonal of 25.4 centimeters (Faytech FT10TMB) was positioned 15 centimeters in front of the animals. The behavioral protocol was implemented using custom MATLAB code and the MATLAB-based software MonkeyLogic (Hwang et al., 2019) to control a data acquisition device (National Instruments PCIe-6323) with a break-out panel (National Instruments BNC-2090A). Time-stamped behavioral event codes were sent to the photometry recording system (see section 5.3.1*)*. Videos of the animals were recorded during the task using MonkeyLogic and an infrared camera (ELP USBFHD04H-BL36IR) pointed at the front of the animal to capture hand and nose movements.

5.2.2 Controlled water protocol

After at least three days of post-operative recovery and three to four weeks of viral expression time, mice were administered a controlled water protocol to motivate them for behavioral experiments. Animals received water during daily training sessions. Mice drank 1000 to 1500 microliters of water per day. If mice drank less than 1000 microliters in a training session, they received additional water from a pipette. Animals were trained every day except for occasional weekend days or holidays. On non-training days animals received 1000 to 1500 microliters of water from a pipette. Body weight and health scores were examined daily to ensure that the body weight was maintained above 80% of the weight prior to the surgery.

5.2.3 Habituation to the experimental setup

Mice received water from a pipette in their home cage during the first three to four days of the controlled water protocol before they were handled by the experimenter. On the first day of habituation the animals were slowly accustomed to the handling by the experimenter and introduced to the body tube for head fixation. To reinforce the habituation, animals received water rewards from a pipette during the handling. On the second or third day of habituation, mice were head-fixed in the setup. To introduce the mice to the licking spout and to measure dopamine release in response to free rewards,

in two sessions the animals received free water rewards from one stationary licking spout. Rewards were dispensed with a uniformly distributed random inter-reward interval of 5 seconds.

5.2.4 Pre-training

In order for mice to learn to respond to two movable licking spouts, they received three to four sessions of spout training, in which only one of the two licking spouts was presented in each trial and a water reward was released upon an instrumental lick. Left and right spouts were presented in pseudo-randomized order. One session lasted until the animals were sated and did not respond anymore. After a maximum of four spout training sessions, when mice consumed at least 800 microliters of water during one session, they were progressed to the full task. Spout training sessions were used for photometry control measurements (see section 5.3.3).

5.2.5 Auditory decision-making task

Mice were trained to perform an instrumental response lick on one of two drinking spouts to obtain a water reward. Water rewards were only dispensed upon a correct lick. The correct spout was indicated by an auditory instruction cue according to different rules (see below). The auditory cue was either low frequency band-pass filtered white noise (4 to 8 kilohertz) or high frequency band-pass filtered white noise (16 to 32 kilohertz). The frequency bands were selected to span two octaves that lie symmetrically within the optimal hearing range of mice (Grothe & Pecka, 2014). The auditory cues were played to the animals at 75-80 decibels sound pressure level. At the beginning of each trial, a gray screen (RGB: 0.1, 0.1, 0.1) on the monitor in front of the animals was switched on to indicate the trial start. The gray screen stayed on during the trial. The monitor was at background luminance between trials. After 1000 milliseconds of gray screen, the auditory cue was played on one of the speakers for 1000 milliseconds. Following the offset of the cue, the two licking spouts were moved into the reach of the animals' tongue and the response window started. In the response window, the first contact of the tongue with either of the two spouts was registered as the behavioral response. When the first lick was on the correct spout, a water reward of 5 microliters was triggered to be dispensed from the target spout. The non-target spout was retracted immediately after a correct lick on the target spout. When the first lick in the response window was on the incorrect spout, both spouts were retracted and an error time out of 4000 milliseconds was initiated in addition to the regular inter-trial interval of 4000 milliseconds. In miss trials, when no lick was detected, both spouts were retracted and the regular inter-trial interval began after the 2000-millisecond response window. The gray screen indicating the active trial status was switched off when the target spout retracted, that is after reward consumption in correct trials, after a false response in incorrect trials, and after no response in miss trials. Trials were presented in blocks of 32 trials in pseudo-randomized order, that is, within a block the conditions were drawn randomly without replacement.

5.2.6 Imaging sessions

To obtain quasi-continuous data points of dopamine signature across the learning process while maximizing data acquisition efficiency, dopamine signals were recorded in every other behavioral session.

5.2.7 Task rules

Animals were trained on three different implicit task rules based on the two instruction cue dimensions. The factors frequency (low or high) and location (left or right speaker) of the sound were fully crossed to yield a total of 4 conditions. Animals in the main experiment were first trained on the location rule: they had to lick the left spout following a sound from the left speaker or the right spout following a sound from the right speaker in order to obtain a reward. Once animals had acquired the first rule, the rule was switched to the previously irrelevant frequency dimension: now animals had to lick the left spout following a low sound and the right spout following a high sound in order to obtain a reward. Finally, after animals had acquired the frequency rule, the rule was switched within the dimension of frequency: animals had to lick the left spout following a high sound or the right spout following a low sound in order to obtain a reward. The order of the task rules was constant in the main experiment, but was varied during pilot experiments (see section 5.2.10).

5.2.8 Session durations and criterion performance

For the definition of criterion performance, sessions were truncated at the end to avoid distortions by experimenter-related variation in session durations and idiosyncratic amounts of error trials accumulated at the end of the session due to satiety and disengagement of the animals. Behavioral sessions were terminated manually when miss trials due to satiety were noticed. Sessions were then truncated post-hoc at the first miss trial after 90 percent of the trials in the behavioral session. Sessions were further truncated if the performance dropped below 1.5 standard deviations during the last 15 percent of trials before the first miss trial. Overall, this method resulted in 5.67 percent of excluded trials. A negligible amount of miss trials remaining in earlier parts of the sessions (0.02 percent of all trials) were also excluded from the analysis. The average session duration after truncating was 329.97 trials (standard deviation 76.99 trials).

Criterion performance for task acquisition was defined as at least 80 percent of correct trials per session and at least 60 percent of correct trials in each of the four single conditions. Task rules were switched after the animals had performed at least two imaging sessions at criterion performance and when the behavior-only session immediately before the switch was also a criterion session. One animal was accidentally switched from the frequency task to the frequency reversed task after one criterion imaging session, but did not show obvious behavioral deviations compared to the other animals. In the final task, the animals were imaged in at least six criterion sessions.

5.2.9 Response bias control

Due to the forced choice between the two spouts, mice naturally developed idiosyncratic preferences for either of the two spouts. Two quantify this preference a response bias index was calculated as *(fraction correct left trials - fraction correct right trials) / (fraction correct left trials)*. A value above 0.5 indicated a preference for the left spout and a value below 0.5 indicated a preference for the right spout. The absolute response bias (|*Response bias*|) presented in figures in the *Results* section was calculated as |*(response bias-0.5)*| and thus ranged between 0 and 0.5 with larger values indicating larger absolute response biases independent of the response side. The response preference could be manipulated by adjusting the relative position of the two licking spouts, that is, positioning the non-preferred spout closer or the preferred spout further away. When extreme side preferences were observed, the spouts were positioned individually for each animal before the session to gently counteract their strong preference in previous sessions. In behavior-only sessions, but not in imaging sessions, the response bias was monitored and spout positions were adjusted online. These intra-session adjustments were only conducted during the acquisition of the first task rule before the animals reached criterion performance, but not in later sessions. The intra-session adjustments were not conducted during imaging sessions to keep the effort of reaching the spouts constant.

5.2.10 Pilot experiments

In pilot experiments, the task rule for the initial task acquisition was varied to examine potential advantages for one or the other instruction cue dimension. The pilot experiment was identical to the main experiment, except for the following difference. Since the licking spouts were separately movable, they could also be used to demonstrate the correct response to the animal. In “no-choice” trials only the spout was presented that would dispense a reward according to the task rule, thereby forcing the animal to make a correct choice. These no-choice trials were used in pilot experiments to speed up task acquisition. In the main experiment, no-choice trials were not used in order to allow straight-forward and rigorous analysis and interpretation of dopamine measurements. In the pilot experiments, animals showed no difference in the initial task acquisition time between location and frequency rules, suggesting that animals did not prefer either one of the instruction cue dimensions per se (see Fig. 5).

5.3 Dopamine fiber photometry

5.3.1 Imaging setup

Fiber photometric signals were acquired with an analog optometer with amplification and filter module (npi electronic FOM-02 and LPBF-01GD). In the fiber optometer, LED excitation light at 470 nanometers was passed through a 442-478 nm excitation filter (Thorlabs). Emitted fluorescence was passed through a 500-530 nanometer emission filter (Thorlabs). Excitation and emission signals were separated using a 495-nanometer dichroic mirror (Thorlabs). For control measurements, green light at 556 nanometers (LED) a 546-566 nanometer excitation filter, a 589-625 nanometer emission filter and a 573-nanometer dichroic mirror were used. The control channel was separated from the dLight channel using a 532-nanometer dichroic mirror.

5.3.2 Dopamine imaging

A low-autofluorescence patch cable (Thorlabs FP400URT-CUSTOM) with FC/PC connector and 1.25-millimeter ferrule ending was connected to the implanted ceramic ferrule using a ceramic mating sleeve (Thorlabs ADAL1). The excitation light intensity was set to 50 microwatts at the tip of the patch cable in all dLight imaging sessions. The transmission rate of the implanted ferrules was between 80 and 86 percent, as tested before implantation. This resulted in an excitation intensity of 40 to 43 microwatts at the site of measurement. The fluorescence signals were amplified and filtered in hardware with a gain of 100 and a low-pass filter at 100 hertz. Both raw and amplified signals were digitized at 1 kilohertz sampling rate using a data acquisition system (Plexon Omniplex) and recorded together with time stamps from the behavioral setup.

5.3.3 Movement control imaging

Control imaging sessions were performed in the same way as dLight imaging sessions, except using the 556-nanometer channel. Since the excitation light in this control channel does not overlap with the excitation wavelength of dLight, the recorded signal is assumed to be background autofluorescence independent of dLight activity. The light intensity was titrated individually for each animal to match the background fluorescence level in the control channel to the level of the regular dLight recordings, resulting in intensities ranging between 60 and 250 microwatts. The background fluorescence of the dLight channel in the control recordings was close to zero. It was therefore assumed that the control recordings contained merely movement artifacts that should be comparable to potential movement artifacts in the dLight recordings (see Fig. 15).

5.4 Data analysis

Data were analyzed with custom Python code, using the packages *NumPy* (Harris et al., 2020), *SciPy* (Virtanen et al., 2020), *pandas* (McKinney, 2010), *scikit-learn* (Pedregosa et al., 2011), and *statsmodels* (Seabold & Perktold, 2010), in addition to toolboxes mentioned below.

5.4.1 Session-based choice model

The session-based choice model was a custom-programed logistic regression model. The probability of an animal's choice $\hat{y}_i$ in each trial was modeled as a linear combination of predictors passed through a logistic function

$$\hat{y} = \frac{1}{1 + e^{-z}}$$

where

$$z = \sum_p \beta_p + \beta_0$$

and where β_p is the regression weight for predictor p and β_0 is an intercept, which represents a general tendency for a left or right response. Regressor weights were optimized by minimizing the negative log-likelihood function

$$J = -\frac{1}{m} \sum_i^m y \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

where m is the number of trials, y_i is the actual choice in trial i, and $\hat{y}_i$ is the predicted probability of choice in trial i. Weights were optimized by gradient descent and the optimization was stopped when the loss was below $1x10^{-4}$ or a maximum number of iterations was reached. Binary model predictions for choice were calculated by rounding the model probability $\hat{y}$ to 0 or 1. The custom model was validated using the Pyglmnet toolbox (Jas et al., 2020), which produced similar results.

The model was fit per session with 10-fold cross-validation. The samples of all regressors were randomly split into training set and test set 10 times, such that every sample appeared in the test set once. The predictions for all samples from the test sets were used to calculate the cross-validated prediction accuracy, which was used for model selection. Using likelihood-based measures such as pseudo-R^2, cross-validated bit per trial (Akrami et al., 2018) or information criteria (Bayesian information criterion, Akaike information criterion) for model selection produced similar results. Prediction accuracy was used for model evaluation due to intuitive interpretation. Elastic net regularization including a grid search for the regularization parameters did not change the results qualitatively. Therefore, no regularization was used.

Regressors for the final model were selected such that the prediction accuracy across all sessions was significantly reduced when a regressor was removed, and the prediction accuracy was not significantly enhanced when other regressors were added or different regressors were used (see Fig. 8). The following regressors were part of the final session-based choice model: Cue location of the current trial (-1 for left speaker, +1 for right speaker), cue frequency of the current trial (-1 for low sound, +1 for high sound), and value difference (the difference of a 10-trial exponentially weighted average rate of rewards up to and including the previous trial calculated separately for each spout,

ranging between -1 and 1). Several alternative models were compared to the final model. The first alternative trial history model included the reward rates of the left and right spout (10-trial exponentially weighted average of rewards up to and including the previous trial, ranging between 0 and 1). The second alternative trial history model included three predictors for the choice up to three trials back (-1 for left choice, +1 for right choice). The third alternative trial history model included two predictors representing a “win-stay” and “lose-switch” strategy, respectively; win-stay was coded as -1 following a correct previous trial with a left choice, as +1 following a correct previous trial with a right choice, and as 0 following a false previous trial; lose-switch was coded as -1 following a false previous trial with a left choice, as +1 following a false previous trial with a right choice, and as 0 following a correct previous trial. The additional variables included the previous choice in the previous trial (-1 for left previous choice, +1 for right previous choice) and previous trial outcome (-1 for no reward in the previous trial, +1 for reward in the previous trial) and the interaction thereof; the previous choice and outcome and interaction thereof up to three trials back; the win-stay and lose-switch predictors; the interaction of the cue predictors; and the history of the cue up to three trials back.

Thus, input variables were all in the range between -1 and +1. Input variables were not standardized to avoid mean-centering and to ascertain interpretability of weight deviations from zero.

5.4.2 Trial-based choice model

The trial-based choice model was fit using the PsyTrack toolbox (Roy et al., 2021), which also models choice in a logistic regression. The PsyTrack model was fit per animal, using optimized hyperparameters to allow for fluctuations of regressor weights throughout the learning process, both across trials and across sessions. The recommended default initial hyperparameters were used. Model predictions were again made with 10-fold cross-validation to calculate the cross-validated prediction accuracy and compare the PsyTrack model to the custom session-based model. The same regressor variables as in the final session-based model were used.

5.4.3 Photometry data

Fiber-photometric signals were acquired at 1000 samples per second. To optimize data handling efficiency while keeping an adequate resolution for dLight signals, raw signals were smoothed with a 50-millisecond running average and down-sampled to 50 samples per second. For analysis of trial-related modulations of dLight, relative fluorescence $\Delta F/F$ was calculated by subtracting a baseline from every sample and dividing it by the baseline using the average amplitude of a 500-millisecond window before the trial start as a baseline. These $\Delta F/F$ signals were normalized using a robust z-score (subtracting the median and dividing by the median absolute deviation) calculated for each session using the analyzed trial sections to account for differences in signal intensities across sessions and animals. Calculating the z-score using only the baseline period or using the whole session did not change the main results.

5.4.4 Neural encoding model

To quantify the contribution of different task variables and movement variables to fluctuations in dopamine, a linear regression model was used to predict the dopamine traces at every timepoint in a trial, similar to approaches previously published (Musall et al., 2019). The timepoints in the trial to be fitted included 2340 milliseconds after the trial start, which included 1000 milliseconds before and after the cue onset and 340 milliseconds for the response, and additional 2340 milliseconds after the response of the animal, which included the reward and reward consumption time. The predicted dopamine was modeled as

$$\hat{y}\,(i,t) \;=\; \sum_{w} \beta_w(t) X_w(i) + \beta_0(t) + \;\varepsilon(i,t)$$

where $\hat{y}\,(i,t)$ is the normalized dLight activity in trial i at timepoint t, $\beta_w(t)$ is the regression weight for whole-trial variable w at timepoint t, $X_w(i)$ is the value of the whole-trial variable w in trial i, $\beta_0(t)$ is the intercept at timepoint t, and $\varepsilon(i,t)$ is the residual variance.

The model was regularized with ridge regularization, applying a penalty to large coefficients to avoid overfitting. Ridge regularization was chosen over Lasso or elastic net regularization because no sparsity constraint was desired, since the goal was not to find sparse predictors, but to calculate the contribution of all predictors. The regularization parameter for ridge regression was set to 1. A grid search for the regularization parameter did not change the main results. The ridge model was fit with least-squares fitting using the SciPy library (Virtanen et al., 2020).

To avoid overfitting, the model was fit with 5-fold cross-validation in order to obtain cross-validated predictions, which were used to quantify the explained variance of the model as the Pearson correlation of the predicted dopamine traces with the actual traces. The trials were split into training set and test set 5 times, such that each trial was part of the test set once. The regressor variables were standardized by subtracting the mean and dividing by the standard deviation. Standardization was performed after the train/test split to avoid leakage of information from the training set to the test set. After each split, the standardization parameters were calculated from the training set and both training and test set were standardized using theses parameters.

The model fitting was implemented using the SciPy library (Virtanen et al., 2020). Each regressor variable was constructed as a matrix (SciPy *sparse matrix*) with the dimensions *number of samples per trial x (number of samples per trial * trials to fit)*, where the data of each regressor in each trial was represented on the diagonal of the submatrix with the dimensions *number of samples per trial x number of samples per trial*. Whole-trial variables were repeated for all samples in the trial. The resulting regressor matrix had the dimensions *(number of regressors * number of samples per trial) x (number of samples per trial * number of trials to fit)*.

The regressors included an intercept, the current trial outcome, previous trial outcome (+1 for a reward in the previous trial, -1 for no reward in the previous trial), congruence to response preference according to stay versus switch trials (+1 for stay trial, -1 for switch trial), congruence to global response preference defined by individual response preference during a given task rule (+1 for preferred trial, -1 for non-preferred trial), congruence to response preference defined by the response bias weight from the trial-based choice model (+1 for trials with response bias weight above 0, -1 for trials with response bias weight below 0), and the instruction cue weight from the trial-based choice

model. The instruction cue weight from the trial-based choice model was adapted so that positive values represented the strategy of following the relevant instruction cue regardless of the task rule (i.e., cue weight values were inverted for the frequency reversed task). Thus, the cue weight regressor used in the encoding model indicated how strongly the animals were pursuing the optimal strategy of following the instruction cue in a given trial. The model was fit separately for each session in order to examine weight trajectories across performance levels and task rules, but a fit per animal yielded similar qualitative results of explained variance and unique contributions.

5.4.5 Body part tracking

Behavioral videos were recorded at 30 frames per second along with the rest of the behavioral data using the MonkeyLogic toolbox (Hwang et al., 2019) that was used for the presentation of the behavioral task. Video resolution was 320 by 240 pixels. The videos were further compressed using the ffmpeg toolbox (http://ffmpeg.org/). The videos showing the animal from the front were analyzed to track body parts across the trial. Left hand, right hand and nose of the animal were tracked using DeepLabCut 2.1.9 (Mathis et al., 2018; Nath et al., 2019). 100 frames were picked randomly from all animals and labeled manually. Additional 100 outlier frames were labeled after the first iteration of training. 95 percent of labeled images were used for training. A ResNet-50-based neural network with default parameters was used. The train error was 0.97 pixels and the test error was 1.25 pixels. The network was then used to analyze videos from all imaging sessions. Visual inspection suggested that the tracking generalized well across sessions. Unreliable frames with suboptimal likelihood of body part detection (likelihood < 0.99) were identified. These unreliable frames amounted to less than 2.5 percent of frames for the left hand, less than 1.3 percent of frames for the right hand, and less than 0.01 percent for the nose. The unreliable frames were replaced with previous neighboring values (or following neighboring values at the beginning of the trial). The original data at 30 frames per second was filtered with a 5-point median filter and up-sampled to 50 samples per second to match the imaging data. For further analysis, X and Y positions of the body parts were baseline-subtracted using the mean position of the first 500 milliseconds after the trial start as a baseline.

5.4.6 Statistical tests

The sample sizes of animal groups for recordings in the different striatal subregions were not determined a priori to have statistical power for tests across individual animals. Instead, the animal numbers were intended to be large enough to avoid idiosyncratic effects and reproduce findings across animals. Indeed, the variability across animals was relatively small (see *Appendix* Fig. A2/A3). Therefore, for tests of statistical significance, sessions from different animals were pooled, when groups of sessions were compared (e.g., comparisons across performance levels). Trials from several sessions were pooled, when different trial types were compared (e.g., comparisons across preferred and non-preferred trials) or for pairwise comparisons of pooled sessions with only one session per animal. The main results were not qualitatively different without pooling (e.g., *Appendix* Fig. A2). Non-parametric tests were used for comparisons across two groups (Wilcoxon signed-rank/rank-sum test) and three or more groups (Kruskal-Wallis test), unless otherwise stated. Statistical tests are specified

in the caption of each individual figure. Significance levels were adjusted for multiple comparisons (Bonferroni correction), as noted in figure captions. Data are presented as mean ± standard error of the mean across animals, sessions, or trials, as specified in each individual figure. Individual results of statistical tests presented in figures are summarized in *Appendix* Table A1,

The time windows for statistical tests of differences between average dopamine transients in trial epochs of instruction cue and outcome (Figs. 16-23) were adjusted to account for differences in dynamics of dopamine signals across striatal subregions, using windows of 1000 milliseconds for the ventral striatum, 800 milliseconds for the dorsomedial striatum, and 600 milliseconds for the dorsolateral striatum. For the trial epoch of the spout movements a general window of 340 milliseconds after the cue offset was used (based on the visualization in Figs. 16-23).

6. Appendix

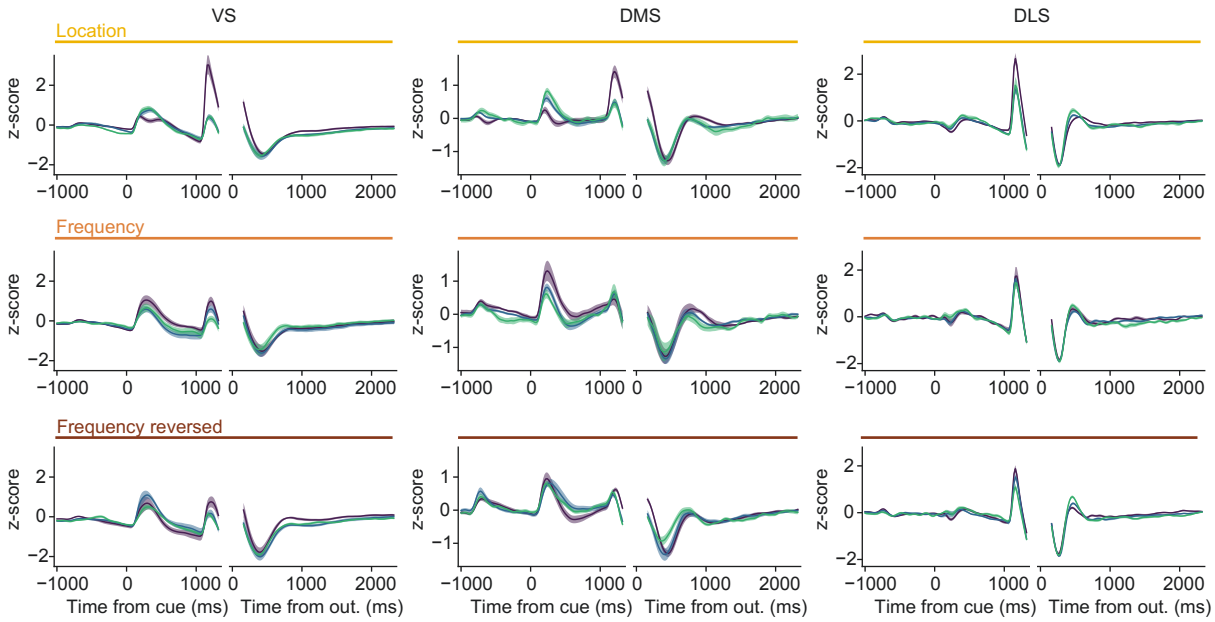


Fig. A1) Average dopamine in error trials across performance levels
Same as Fig. 17A, but error trials instead of correct trials.

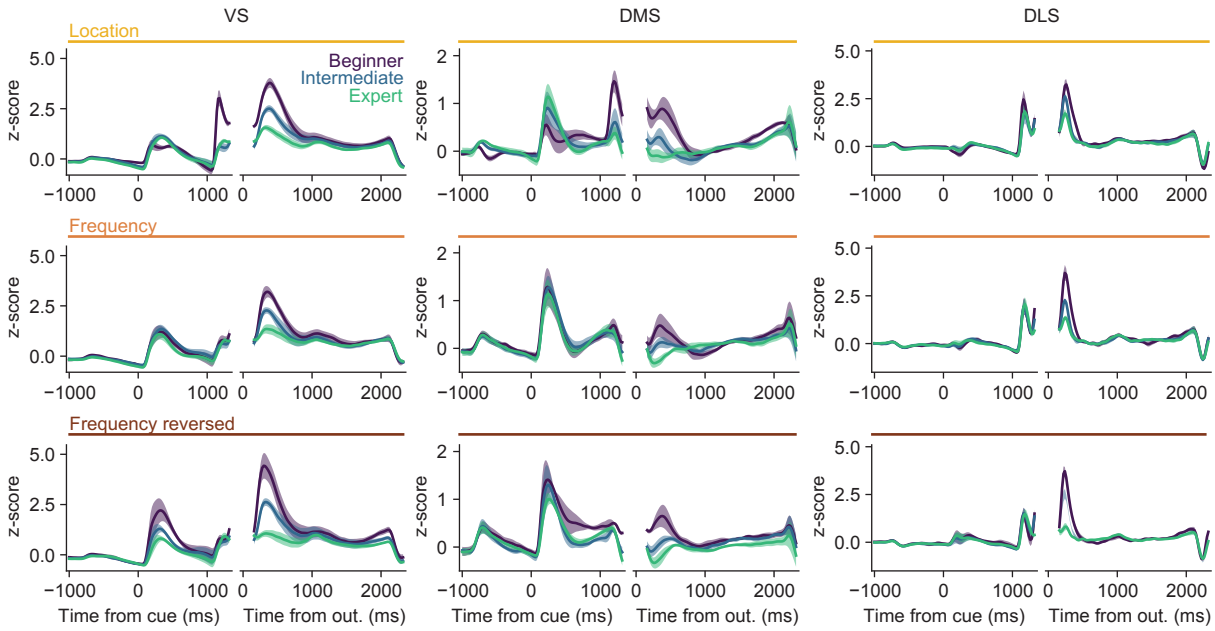


Fig. A2) Average dopamine across performance levels
Same as Fig. 17A, but mean and standard error of the mean calculated across animals instead of pooled sessions.

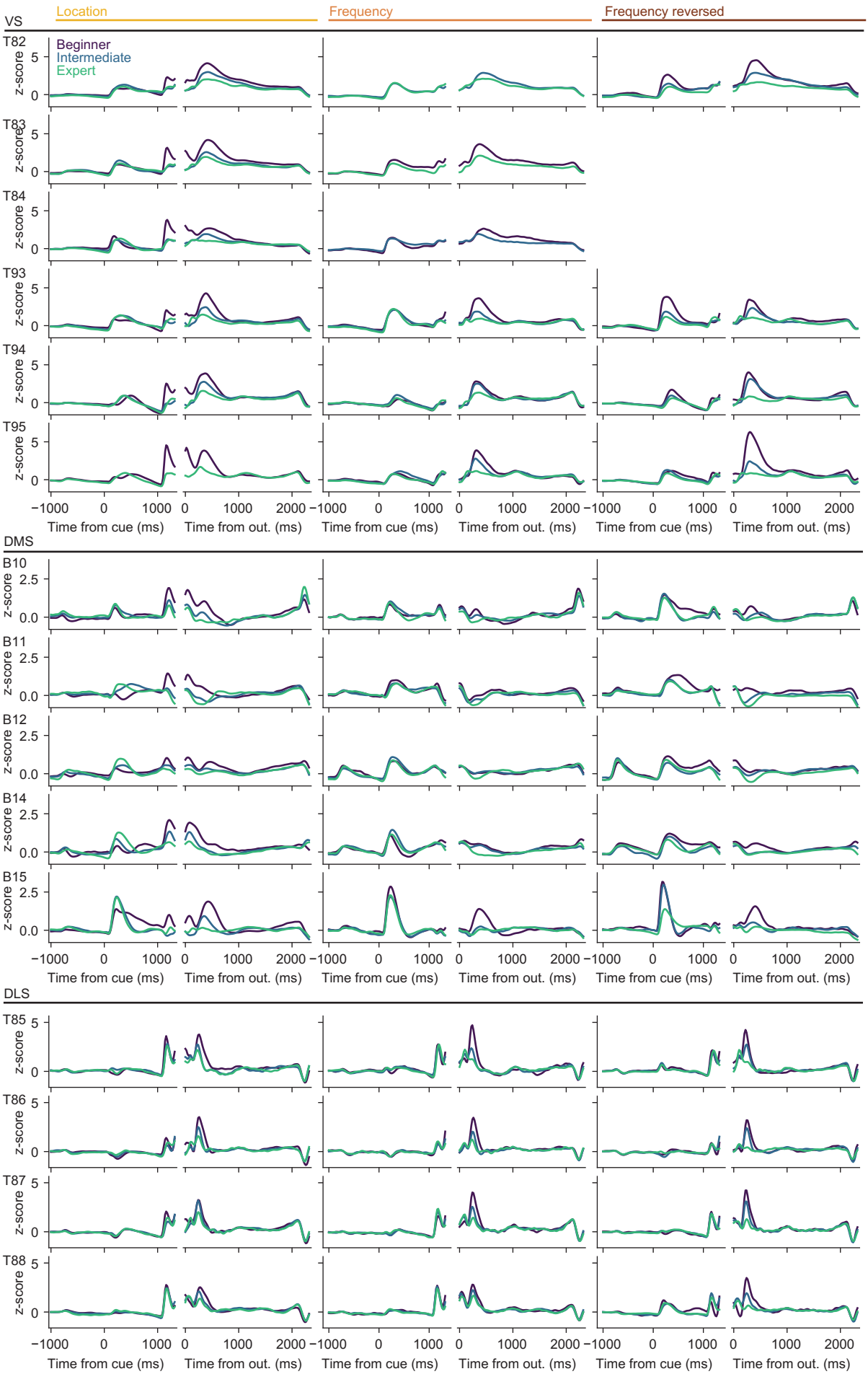
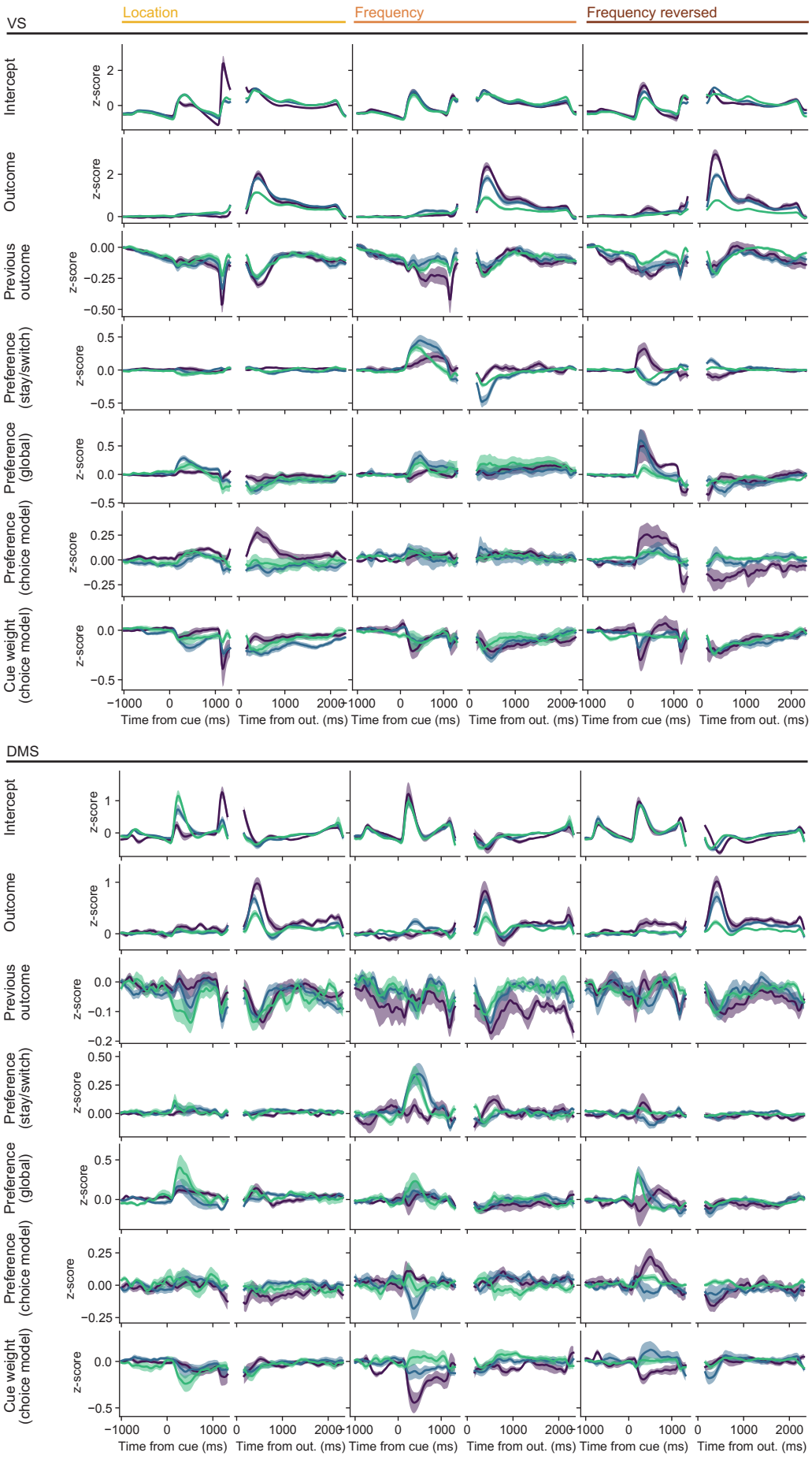

Location
Frequency
Frequency reversed
VS
DMS
DLS
Beginner
Intermediate
Expert
T82
T83
T84
T93
T94
T95
B10
B11
B12
B14
B15
T85
T86
T87
T88
z-score
Time from cue (ms)
Time from out. (ms)

Fig. A3) Average dopamine across performance levels per animal
Same as Fig. 17A, but plotted for each animal separately. Lines show mean across trials.


VS
Location
Frequency
Frequency reversed
Intercept
Outcome
Previous outcome
Preference (stay/switch)
Preference (global)
Preference (choice model)
Cue weight (choice model)
z-score
Time from cue (ms)
Time from out. (ms)
DMS

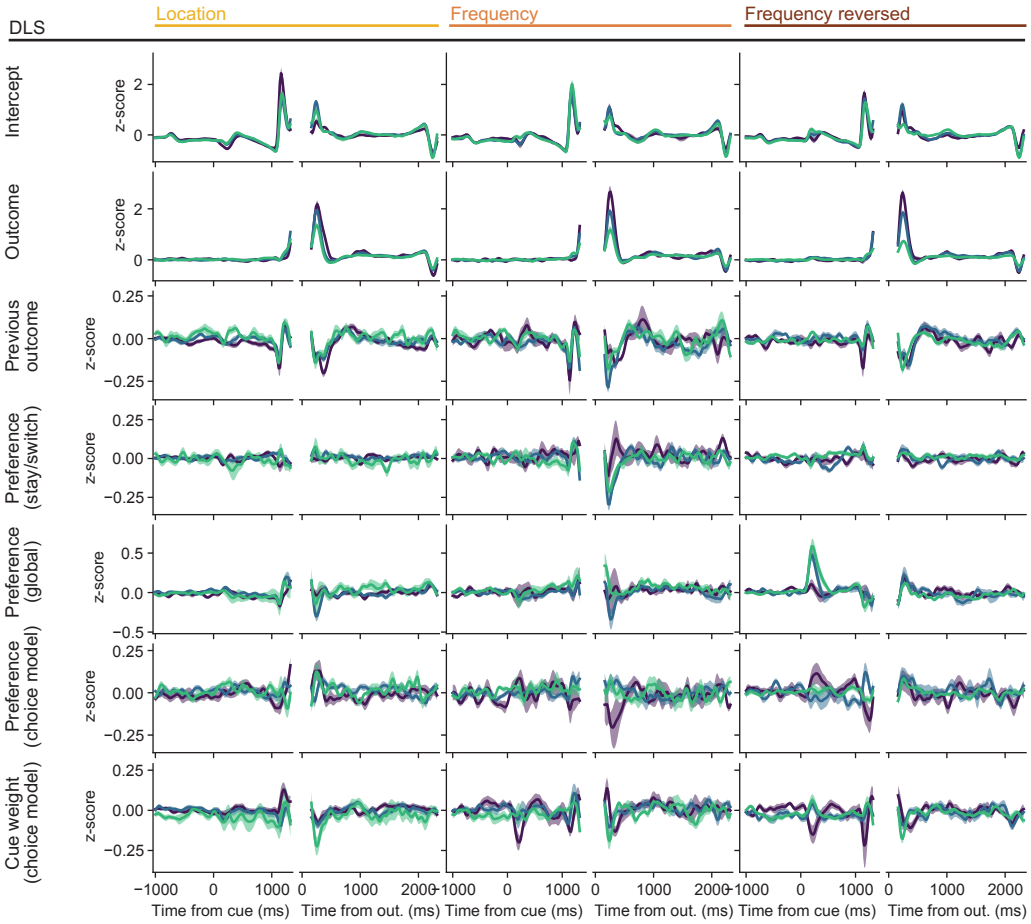


Fig. A4) Encoding model weights (all)
Normalized regressor weights across the trial for all regressors in the full encoding model (see Fig. 24 and *Methods*) split by performance levels and task rules. Mean ± standard error of the mean across sessions.

Table A1) Statistical tests in figures

KW = Kruskal-Wallis H-test; PCS = Pearson's chi-squared test; PT = Permutation test; WRS = Wilcoxon rank-sum test; WSR = Wilcoxon signed-rank test; n = sample size (per group, if applicable); p = p value (after correction for multiple comparisons, if applicable)

Figure	**Comparison** (see figure caption and main text for details)	**Test**	**n**	**p**
4E	Location vs. Frequency	WSR	24	1.1921e-07
4E	Frequency vs. Frequency reversed	WSR	24	0.0002
4E	Location vs. Frequency reversed	WSR	24	1.1921e-07
5B	Location vs. Frequency	PT	4, 4	0.8857
8B	Full vs. w/o Cue freq.	WSR	692	3.4722e-70
8B	Full vs. w/o Cue loc.	WSR	692	6.7205e-42
8B	Full vs. w/o Value diff.	WSR	692	1.7629e-23
8B	Full vs. Rew. rate instead	WSR	692	1.0000
8B	Full vs. Choice hist. instead	WSR	692	1.0000
8B	Full vs. WSLS instead	WSR	692	2.0034e-05
8B	Full vs. + Choice/rew.	WSR	692	1.0000
8B	Full vs. + Choice/rew. hist.	WSR	692	1.0000
8B	Full vs. + WSLS	WSR	692	1.0000
8B	Full vs. + WSLS hist.	WSR	692	1.0000
8B	Full vs. + Cue interaction	WSR	692	1.0000
8B	Full vs. + Cue hist.	WSR	692	0.0070
8C	Bias vs. History (All)	WSR	692	7.6469e-15
8C	History vs. Cue (All)	WSR	692	1.9293e-79
8C	Cue vs. Full (All)	WSR	692	6.3491e-24
8C	Bias vs. History (Beginner)	WSR	193	6.1335e-10
8C	History vs. Cue (Beginner)	WSR	193	0.2288
8C	Cue vs. Full (Beginner)	WSR	193	1.2032e-12
8C	Bias vs. History (Intermediate)	WSR	190	1.0518e-09
8C	History vs. Cue (Intermediate)	WSR	190	3.6535e-24
8C	Cue vs. Full (Intermediate)	WSR	190	1.8029e-11
8C	Bias vs. History (Expert)	WSR	309	0.1196
8C	History vs. Cue (Expert)	WSR	309	2.2566e-51
8C	Cue vs. Full (Expert)	WSR	309	1.0000
10A	Bias vs. History (All)	WSR	692	3.5470e-014
10A	History vs. Cue (All)	WSR	692	5.7668e-101
10A	Cue vs. Full (All)	WSR	692	0.0635
10A	Bias vs. History (Beginner)	WSR	193	1.0000
10A	History vs. Cue (Beginner)	WSR	193	2.5319e-011
10A	Cue vs. Full (Beginner)	WSR	193	1.0000
10A	Bias vs. History (Intermediate)	WSR	190	1.0000
10A	History vs. Cue (Intermediate)	WSR	190	2.7364e-031
10A	Cue vs. Full (Intermediate)	WSR	190	0.7947
10A	Bias vs. History (Expert)	WSR	309	2.7548e-019
10A	History vs. Cue (Expert)	WSR	309	2.2565e-051
10A	Cue vs. Full (Expert)	WSR	309	0.2286
10B	Trial-based cue vs. Sess.-based full (All)	WSR	692	2.0412e-19
10B	Trial-based cue vs. Sess.-based full (Beginner)	WSR	193	5.8026e-12
10B	Trial-based cue vs. Sess.-based full (Intermediate)	WSR	309	3.0066e-11
10B	Trial-based cue vs. Sess.-based full (Expert)	WSR	190	0.6193
10B	Trial-based cue vs. Sess.-based full (First freq.)	WSR	24	1.7811e-03
10B	Trial-based cue vs. Sess.-based full (First freg. rev.)	WSR	24	3.9339e-05
10C	Trial-based cu vs. + Value diff.	WSR	692	0.0318
10C	Trial-based cue vs. + Prev. choice	WSR	692	0.6471
10C	Trial-based cue vs. + Rew. rate L/R	WSR	692	0.6471
16B	First sessions (Cue epoch, VS)	KW	6, 6, 6	0.0691
16B	First sessions (Spouts epoch, VS)	KW	6, 6, 6	0.0021
16B	First sessions (Out. epoch, VS)	KW	6, 6, 6	0.7000
16B	First sessions (Cue epoch, DMS)	KW	5, 5, 5	0.0344
16B	First sessions (Spouts epoch, DMS)	KW	5, 5, 5	0.0077
16B	First sessions (Out. epoch, DMS)	KW	5, 5, 5	0.4025
16B	First sessions (Cue epoch, DLS)	KW	4, 4, 4	0.5004
16B	First sessions (Spouts epoch, DLS)	KW	4, 4, 4	0.2457
16B	First sessions (Out. epoch, DLS)	KW	4, 4, 4	0.0775
16D	Pre-training, first sessions (Spout epoch, VS)	KW	6, 6	0.0163

16D	Pre-training, first sessions (Out. epoch, VS)	KW	6, 6	0.0039
16D	Pre-training, first sessions (Spout epoch, DMS)	KW	5, 5	0.0031
16D	Pre-training, first sessions (Out. epoch, DMS)	KW	5, 5	0.0226
16D	Pre-training, first sessions (Spout epoch, DLS)	KW	4, 4	0.0833
16D	Pre-training, first sessions (Out. epoch, DLS)	KW	4, 4	0.0209
17B	Performance levels (Cue epoch, VS, Loc.)	KW	25, 15, 16	0.8692
17B	Performance levels (Cue epoch, VS, Freq.)	KW	6, 9, 12	0.0669
17B	Performance levels (Cue epoch, VS, Freq. rev.)	KW	6, 9, 34	0.0209
17B	Performance levels (Cue epoch, DMS, Loc.)	KW	21, 22, 12	0.0315
17B	Performance levels (Cue epoch, DMS, Freq.)	KW	5, 12, 10	0.2699
17B	Performance levels (Cue epoch, DMS, Freq. rev.)	KW	10, 11, 40	4.7073e-05
17B	Performance levels (Cue epoch, DLS, Loc.)	KW	26, 24, 8	0.1356
17B	Performance levels (Cue epoch, DLS, Freq.)	KW	5, 11, 8	0.1465
17B	Performance levels (Cue epoch, DLS, Freq. rev.)	KW	11, 14, 33	0.5401
17C	Performance levels (Spout epoch, VS, Loc.)	KW	25, 15, 16	0.0005
17C	Performance levels (Spout epoch, VS, Freq.)	KW	6, 9, 12	0.6343
17C	Performance levels (Spout epoch, VS. Freq. rev.)	KW	6, 9, 34	0.7745
17C	Performance levels (Spout epoch, DMS, Loc.)	KW	21, 22, 12	8.2775e-05
17C	Performance levels (Spout epoch, DMS, Freq.)	KW	5, 12, 10	0.5546
17C	Performance levels (Spout epoch, DMS. Freq. rev.)	KW	10, 11, 40	0.0020
17C	Performance levels (Spout epoch, DLS, Loc.)	KW	26, 24, 8	0.0046
17C	Performance levels (Spout epoch, DLS, Freq.)	KW	5, 11, 8	0.8281
17C	Performance levels (Spout epoch, DLS. Freq. rev.)	KW	11, 14, 33	0.2653
17D	Performance levels (Out. epoch, VS, Loc.)	KW	25, 15, 16	1.9180e-08
17D	Performance levels (Out. epoch, VS, Freq.)	KW	6, 9, 12	0.0042
17D	Performance levels (Out. epoch, VS. Freq. rev.)	KW	6, 9, 34	0.0001
17D	Performance levels (Out. epoch, DMS, Loc.)	KW	21, 22, 12	6.2314e-06
17D	Performance levels (Out. epoch, DMS, Freq.)	KW	5, 12, 10	0.0331
17D	Performance levels (Out. epoch, DMS. Freq. rev.)	KW	10, 11, 40	1.9639e-06
17D	Performance levels (Out. epoch, DLS, Loc.)	KW	26, 24, 8	1.7177e-08
17D	Performance levels (Out. epoch, DLS, Freq.)	KW	5, 11, 8	0.0009
17D	Performance levels (Out. epoch, DLS. Freq. rev.)	KW	11, 14, 33	5.8512e-10
18B	Last loc. vs First 10% freq. (Cue epoch, VS)	WRS	1479, 129	1.3805e-05
18B	Last loc. vs First freq. (Cue epoch, VS)	WRS	1479, 1297	1.3655e-07
18B	First 10% freq. vs First freq. (Cue epoch, VS)	WRS	129, 1297	0.0996
18B	Last loc. vs First 10% freq. (Out. epoch, VS)	WRS	1479, 129	1.1493e-18
18B	Last loc. vs First freq. (Out. epoch, VS)	WRS	1479, 1297	2.3253e-90
18B	First 10% freq. vs First freq. (Out. epoch, VS)	WRS	129, 1297	1.0000
18B	Last loc. vs First 10% freq. (Cue epoch, DMS)	WRS	1127, 90	0.0099
18B	Last loc. vs First freq. (Cue epoch, DMS)	WRS	1127, 1011	0.5261
18B	First 10% freq. vs First freq. (Cue epoch, DMS)	WRS	90, 1011	0.0431
18B	Last loc. vs First 10% freq. (Out. epoch, DMS)	WRS	1127, 90	5.3109e-05
18B	Last loc. vs First freq. (Out. epoch, DMS)	WRS	1127, 1011	4.8133e-10
18B	First 10% freq. vs First freq. (Out. epoch, DMS)	WRS	90, 1011	0.2245
18B	Last loc. vs First 10% freq. (Cue epoch, DLS)	WRS	991, 89	1.0000
18B	Last loc. vs First freq. (Cue epoch, DLS)	WRS	991, 860	1.0000
18B	First 10% freq. vs First freq. (Cue epoch, DLS)	WRS	89, 860	0.8843
18B	Last loc. vs First 10% freq. (Out. epoch, DLS)	WRS	991, 89	8.7062e-06
18B	Last loc. vs First freq. (Out. epoch, DLS)	WRS	991, 860	1.8500e-44
18B	First 10% freq. vs First freq. (Out. epoch, DLS)	WRS	89, 860	0.8321
18D	Last freq. vs First 10% freq. rev. (Cue epoch, VS)	WRS	1695, 37	0.0358
18D	Last freq. vs First freq. rev. (Cue epoch, VS)	WRS	1695, 866	7.9046e-019
18D	First 10% freq. rev. vs. First freq. (Out. epoch, VS)	WRS	37, 866	1.8025e-004
18D	Last freq. vs First 10% freq. rev. (Out. epoch, VS)	WRS	1695, 37	3.7833e-013
18D	Last freq. vs First freq. rev. (Out. epoch, VS)	WRS	1695, 866	6.4963e-145
18D	First 10% freq. rev. vs. First freq. (Cue epoch, VS)	WRS	37, 866	0.6403
18D	Last freq. vs First 10% freq. rev. (Cue epoch, DMS)	WRS	1207, 34	0.6460
18D	Last freq. vs First freq. rev. (Cue epoch, DMS)	WRS	1207, 954	2.4691e-31
18D	First 10% freq. rev. vs. First freq. (Out. epoch, DMS)	WRS	34, 954	0.3335
18D	Last freq. vs First 10% freq. rev. (Out. epoch, DMS)	WRS	1207, 34	1.3273e-07
18D	Last freq. vs First freq. rev. (Out. epoch, DMS)	WRS	1207, 954	2.8932e-27
18D	First 10% freq. rev. vs. First freq. (Cue epoch, DMS)	WRS	34, 954	0.0080
18D	Last freq. vs First 10% freq. rev. (Cue epoch, DLS)	WRS	1025, 44	0.0466
18D	Last freq. vs First freq. rev. (Cue epoch, DLS)	WRS	1025, 689	0.7010
18D	First 10% freq. rev. vs. First freq. (Out. epoch, DLS)	WRS	44, 689	0.0267
18D	Last freq. vs First 10% freq. rev. (Out. epoch, DLS)	WRS	1025, 44	1.3007e-07
18D	Last freq. vs First freq. rev. (Out. epoch, DLS)	WRS	1025, 689	2.7527e-59
18D	First 10% freq. rev. vs. First freq. (Cue epoch, DLS)	WRS	44, 689	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, L-1, VS)	WRS	755, 724	0.7267
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F0, VS)	WRS	1038, 259	1.3397e-10
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F1, VS)	WRS	1043, 580	1.0503e-46
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F2, VS)	WRS	829, 611	8.4611e-28

19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, L-1, VS)	WRS	755, 724	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F0, VS)	WRS	1038, 259	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F1, VS)	WRS	1043, 580	2.8480e-19
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F2, VS)	WRS	829, 611	4.7271e-19
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, L-1, VS)	PCS	755 724; 89, 118	0.1443
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F0, VS)	PCS	1038, 259; 143, 927	1.7260e-227
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F1, VS)	PCS	1043, 580, 180, 652	5.6521e-88
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F2, VS)	PCS	829, 611; 63, 278	1.3413e-37
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, L-1, DMS)	WRS	572, 555	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F0, DMS)	WRS	829, 182	0.2914
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F1, DMS)	WRS	677, 345	0.0021
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F2, DMS)	WRS	714, 490	1.0450e-19
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, L-1, DMS)	WRS	572, 555	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F0, DMS)	WRS	829, 182	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F1, DMS)	WRS	677, 345	0.3198
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F2, DMS)	WRS	714, 490	0.0230
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, L-1, DMS)	PCS	572, 555; 76, 92	0.8434
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F0, DMS)	PCS	829, 182; 143, 791	1.2020e-188
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F1, DMS)	PCS	677, 345; 82, 419	1.0804e-73
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F2, DMS)	PCS	714, 490; 78, 297	8.0124e-38
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, L-1, DLS)	WRS	488, 503	0.0039
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F0, DLS)	WRS	651, 209	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F1, DLS)	WRS	622, 404	0.3704
19B	Pref. (stay) vs. non.pref. (switch) (Cue epoch, F2, DLS)	WRS	593, 416	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, L-1, DLS)	WRS	488, 503	0.7366
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F0, DLS)	WRS	651, 209	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Out. epoch, F1, DLS)	WRS	622, 404	0.0512
19B	Pref. (stay) vs. non.pref. (switch) (Out. Epoch, F2, DLS)	WRS	593, 416	0.0025
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, L-1, DLS)	PCS	488, 503; 92, 85	1.0000
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F0, DLS)	PCS	651, 209; 146, 592	6.1004e-109
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F1, DLS)	PCS	622, 404; 131, 350	1.0003e-32
19B	Pref. (stay) vs. non.pref. (switch) (Behavior, F2, DLS)	PCS	593, 416; 85, 257	1.7338e-26
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, B., VS)	WRS	1037, 296	2.8355e-09
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, I., VS)	WRS	1437, 802	1.4306e-55
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, E., VS)	WRS	1564, 1345	3.1531e-35
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, B., VS)	WRS	1037, 296	0.1000
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, I., VS)	WRS	1437, 802	4.6608e-19
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, E., VS)	WRS	1564, 1345	2.6345e-09
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, B., VS)	PCS	1037, 296; 199, 940	1.9970e-195
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, I., VS)	PCS	1437, 802; 151, 796	8.0578e-136
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, E., VS)	PCS	1564, 1345; 90, 314	1.0688e-31
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, B., DMS)	WRS	829, 182	0.2186
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, I., DMS)	WRS	1722, 1007	1.0827e-35
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, E., DMS)	WRS	1294, 1112	3.4373e-13
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, B., DMS)	WRS	829, 182	1.0000
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, I., DMS)	WRS	1722, 1007	0.0022
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, E., DMS)	WRS	1294, 1112	1.0000
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, B., DMS)	PCS	829, 182; 143, 791	9.0149e-189
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, I., DMS)	PCS	1722, 1007; 195, 898	3.2548e-140
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, E., DMS)	PCS	1294, 1112; 62, 239	1.0982e-26
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, B., DLS)	WRS	781, 269	1.0000
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, I., DLS)	WRS	1536, 1059	1.0000
20B	Pref. (stay) vs. non-pref. (switch) (Cue epoch, E., DLS)	WRS	1054, 902	0.9339
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, B., DLS)	WRS	781, 269	1.0000
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, I., DLS)	WRS	1536, 1059	1.0540e-09
20B	Pref. (stay) vs. non-pref. (switch) (Out. epoch, E., DLS)	WRS	1054, 902	1.1991e-03
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, B., DLS)	PCS	781, 269; 179, 698	3.7234e-122
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, I., DLS)	PCS	1536, 1059; 245, 715	2.6093e-70
20B	Pref. (stay) vs. non-pref. (switch) (Behavior, E., DLS)	PCS	1054, 902; 97, 253	9.6838e-19
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, B., VS)	WRS	3327, 1608	0.0673
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, I., VS)	WRS	1886, 1639	4.9303e-18
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, E., VS)	WRS	1405, 140	7.9003e-12
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, B., VS)	WRS	3327, 1608	4.3517e-06
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, I., VS)	WRS	1886, 1639	1.0535e-74
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, E., VS)	WRS	1405, 140	4.2704e-43
21B	Pref. (global) vs. non-pref. (global) (Behavior, B., VS)	PCS	3327, 1608; 1055, 3246	0.0000
21B	Pref. (global) vs. non-pref. (global) (Behavior, I., VS)	PCS	1886, 1639; 618. 868	5.0039e-14
21B	Pref. (global) vs. non-pref. (global) (Behavior, E., VS)	PCS	1405, 140; 244, 244	1.0000
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, B., DMS)	WRS	3160, 907	5.9595e-20
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, I., DMS)	WRS	2154, 2576	7.8242e-20
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, E., DMS)	WRS	1206, 1115	2.7628e-06
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, B., DMS)	WRS	3160, 907	0.3980
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, I., DMS)	WRS	2154, 2576	4.9151e-20
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, E., DMS)	WRS	1206, 1115	2.9712e-05

21B	Pref. (global) vs. non-pref. (global) (Behavior, B., DMS)	PCS	3160, 907; 617, 3051	0.0000
21B	Pref. (global) vs. non-pref. (global) (Behavior, I., DMS)	PCS	2154, 2576; 1265, 866	1.2289e-25
21B	Pref. (global) vs. non-pref. (global) (Behavior, E., DMS)	PCS	1206, 1115; 156, 255	6.6508e-07
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, B., DLS)	WRS	3538, 1612	0.0187
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, I., DLS)	WRS	2894, 2622	6.0066e-04
21B	Pref. (global) vs. non-pref. (global) (Cue epoch, E., DLS)	WRS	1019, 961	1.0000
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, B., DLS)	WRS	3538, 1612	0.7924
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, I., DLS)	WRS	2894, 2622	4.3551e-07
21B	Pref. (global) vs. non-pref. (global) (Out. epoch, E., DLS)	WRS	1019, 961	0.4592
21B	Pref. (global) vs. non-pref. (global) (Behavior, B., DLS)	PCS	3538, 1612; 1491, 3423	0.0000
21B	Pref. (global) vs. non-pref. (global) (Behavior, I., DLS)	PCS	2894, 2622; 869, 1148	2.0164e-12
21B	Pref. (global) vs. non-pref. (global) (Behavior, E., DLS)	PCS	1019, 961; 165, 226	0.0030
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, B., VS)	WRS	1267, 83	1.4733e-20
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, I., VS)	WRS	1014, 1252	1.1623e-74
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, E., VS)	WRS	988, 893	1.0368e-08
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, B., VS)	WRS	1267, 83	1.0000
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, I., VS)	WRS	1014, 1252	3.8953e-23
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, E., VS)	WRS	988, 893	1.0000
22B	Pref. (global) vs. non-pref. (global) (Behavior, B., VS)	PCS	1267, 83; 234. 1413	0.0000
22B	Pref. (global) vs. non-pref. (global) (Behavior, I., VS)	PCS	1014, 1252; 622, 379	2.1303e-19
22B	Pref. (global) vs. non-pref. (global) (Behavior, E., VS)	PCS	988, 893; 98, 200	1.3335e-09
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, B., DMS)	WRS	1636, 483	0.0260
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, I., DMS)	WRS	1457, 1491	1.4256e-15
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, E., DMS)	WRS	1361, 1272	1.6715e-09
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, B., DMS)	WRS	1636, 483	0.0023
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, I., DMS)	WRS	1457, 1491	1.2974e-11
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, E., DMS)	WRS	1361, 1272	3.1513e-05
22B	Pref. (global) vs. non-pref. (global) (Behavior, B., DMS)	PCS	1636, 483; 743, 1901	8.2197e-248
22B	Pref. (global) vs. non-pref. (global) (Behavior, I., DMS)	PCS	1457, 1491; 645, 608	0.7094
22B	Pref. (global) vs. non-pref. (global) (Behavior, E., DMS)	PCS	1361, 1272; 186, 266	1.2979e-004
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, B., DLS)	WRS	1437, 720	0.3430
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, I., DLS)	WRS	1634, 1638	5.7009e-33
22B	Pref. (global) vs. non-pref. (global) (Cue epoch, E., DLS)	WRS	965, 956	4.3877e-24
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, B., DLS)	WRS	1437, 720	1.0000
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, I., DLS)	WRS	1634, 1638	1.0000
22B	Pref. (global) vs. non-pref. (global) (Out. epoch, E., DLS)	WRS	965, 956	1.0000
22B	Pref. (global) vs. non-pref. (global) (Behavior, B., DLS)	PCS	1437, 720; 765, 1486	2.6648e-103
22B	Pref. (global) vs. non-pref. (global) (Behavior, I., DLS)	PCS	1634, 1638; 629, 632	1.0000
22B	Pref. (global) vs. non-pref. (global) (Behavior, E., DLS)	PCS	965, 956; 157, 165	1.0000
23B	Prev. corr. vs. prev. false (Cue epoch, B., VS)	WRS	3782, 3820	9.7999e-32
23B	Prev. corr. vs. prev. false (Cue epoch, I., VS)	WRS	5612, 2405	2.1836e-35
23B	Prev. corr. vs. prev. false (Cue epoch, E., VS)	WRS	6623, 952	4.2845e-43
23B	Prev. corr. vs. prev. false (Out. epoch, B., VS)	WRS	3782, 3820	4.4233e-26
23B	Prev. corr. vs. prev. false (Out. epoch, I., VS)	WRS	5612, 2405	1.5554e-42
23B	Prev. corr. vs. prev. false (Out. epoch, E., VS)	WRS	6623, 952	1.2392e-34
23B	Prev. corr. vs. prev. false (Behavior, B., VS)	PCS	3782, 3820; 3820, 3712	0.7222
23B	Prev. corr. vs. prev. false (Behavior, I., VS)	PCS	5612, 2405; 2404, 1010	1.0000
23B	Prev. corr. vs. prev. false (Behavior, E., VS)	PCS	6623, 952; 956, 224	7.3170e-09
23B	Prev. corr. vs. prev. false (Cue epoch, B., DMS)	WRS	3532, 3647	0.0095
23B	Prev. corr. vs. prev. false (Cue epoch, I., DMS)	WRS	7353, 3025	0.0545
23B	Prev. corr. vs. prev. false (Cue epoch, E., DMS)	WRS	6423, 918	1.8308e-06
23B	Prev. corr. vs. prev. false (Out. epoch, B., DMS)	WRS	3532, 3647	1.2620e-20
23B	Prev. corr. vs. prev. false (Out. epoch, I., DMS)	WRS	7353, 3025	3.3327e-10
23B	Prev. corr. vs. prev. false (Out. epoch, E., DMS)	WRS	6423, 918	4.5715e-11
23B	Prev. corr. vs. prev. false (Behavior, B., DMS)	PCS	3532, 3647; 3652, 3765	1.0000
23B	Prev. corr. vs. prev. false (Behavior, I., DMS)	PCS	7353, 3025; 3029, 1429	0.0013
23B	Prev. corr. vs. prev. false (Behavior, E., DMS)	PCS	6423, 918; 919, 234	2.8900e-12
23B	Prev. corr. vs. prev. false (Cue epoch, B., DLS)	WRS	4205, 4136	0.4696
23B	Prev. corr. vs. prev. false (Cue epoch, I., DLS)	WRS	8382, 2971	1.0000
23B	Prev. corr. vs. prev. false (Cue epoch, E., DLS)	WRS	4992, 853	0.9534
23B	Prev. corr. vs. prev. false (Out. epoch, B., DLS)	WRS	4205, 4136	4.1131e-28
23B	Prev. corr. vs. prev. false (Out. epoch, I., DLS)	WRS	8382, 2971	2.4076e-24
23B	Prev. corr. vs. prev. false (Out. epoch, E., DLS)	WRS	4992, 853	5.2266e-09
23B	Prev. corr. vs. prev. false (Behavior, B., DLS)	PCS	4205, 4136; 4137, 3876	0.3721
23B	Prev. corr. vs. prev. false (Behavior, I., DLS)	PCS	8382, 2971; 2969, 1249	5.7598e-05
23B	Prev. corr. vs. prev. false (Behavior, E., DLS)	PCS	4992, 853; 853, 199	0.0019
24B	Shuffle vs. Full (VS)	WSR	132	2.0909e-23
24B	Shuffle vs. Full (DMS)	WSR	143	3.2498e-25
24B	Shuffle vs. Full (DLS)	WSR	140	1.0115e-24
25C	Outcome (VS)	WSR	132	1.5061e-22
25C	Previous outcome (VS)	WSR	132	2.8544e-15
25C	Cue weight (choice model) (VS)	WSR	132	2.9798e-20

25C	Preference (choice model) (VS)	WSR	132	2.9932e-05
25C	Preference (stay/switch) (VS)	WSR	132	4.5667e-07
25C	Preference (global) (VS)	WSR	132	4.0691e-18
25C	Outcome (DMS)	WSR	143	1.3398e-20
25C	Previous outcome (DMS)	WSR	143	5.4112e-05
25C	Cue weight (choice model) (DMS)	WSR	143	2.7069e-21
25C	Preference (choice model) (DMS)	WSR	143	1.0000
25C	Preference (stay/switch) (DMS)	WSR	143	5.9411e-05
25C	Preference (global) (DMS)	WSR	143	4.5851e-13
25C	Outcome (DLS)	WSR	140	8.3807e-24
25C	Previous outcome (DLS)	WSR	140	8.6195e-07
25C	Cue weight (choice model) (DLS)	WSR	140	5.8487e-16
25C	Preference (choice model) (DLS)	WSR	140	0.0043
25C	Preference (stay/switch) (DLS)	WSR	140	0.0003
25C	Preference (global) (DLS)	WSR	140	8.8094e-21

7. References

Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M., & Gardner, J. L. (2016). Adaptable history biases in human perceptual decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(25), E3548-3557. https://doi.org/10.1073/pnas.1518786113

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *34B*, 77-98. https://doi.org/10.1080/14640748208400878

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *33*(2b), 109-121. https://doi.org/10.1080/14640748108400816

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., & Ribas, R. (2019). Solving Rubik's cube with a robot hand. *arXiv*, *preprint arXiv:1910.07113*.

Akrami, A., Kopec, C. D., Diamond, M. E., & Brody, C. D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, *554*(7692), 368-372. https://doi.org/10.1038/nature25510

Amo, R., Yamanaka, A., Tanaka, K. F., Uchida, N., & Watabe-Uchida, M. (2020). A gradual backward shift of dopamine responses during associative learning. *bioRxiv*. https://doi.org/10.1101/2020.10.04.325324

Babayan, B. M., Uchida, N., & Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature Communications*, *9*(1), 1891. https://doi.org/10.1038/s41467-018-04397-0

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407-419. https://doi.org/10.1016/s0028-3908(98)00033-1

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48-69. https://doi.org/10.1038/npp.2009.131

Balsters, J. H., Zerbi, V., Sallet, J., Wenderoth, N., & Mars, R. B. (2020). Primate homologs of mouse cortico-striatal circuits. *eLife*, *9*. https://doi.org/10.7554/eLife.53680

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, *13*, 834-846.

Bayer, H. M., Lau, B., & Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of Neurophysiology*, *98*(3), 1428-1439. https://doi.org/10.1152/jn.01140.2006

Beier, K. T., Steinberg, E. E., DeLoach, K. E., Xie, S., Miyamichi, K., Schwarz, L., Gao, X. J., Kremer, E. J., Malenka, R. C., & Luo, L. (2015). Circuit architecture of VTA dopamine neurons revealed by systematic input-output mapping. *Cell*, *162*(3), 622-634. https://doi.org/10.1016/j.cell.2015.07.015

Belujon, P., & Grace, A. A. (2017). Dopamine system dysregulation in major depressive disorders. *International Journal of Neuropsychopharmacology*, *20*(12), 1036-1046. https://doi.org/10.1093/ijnp/pyx056

Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, *21*(6), 787-793. https://doi.org/10.1038/s41593-018-0152-y

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, *35*(7), 1124-1143. https://doi.org/10.1111/j.1460-9568.2012.07990.x

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research reviews*, *28*(3), 309-369. https://doi.org/10.1016/S0165-0173(98)00019-8

Birkmayer, W., & Hornykiewicz, O. (1961). The L-3,4-dioxyphenylalanine (DOPA)-effect in Parkinson-akinesia. *Wiener klinische Wochenschrift*, *73*, 787-788.

Björklund, A., & Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends in Neurosciences*, *30*(5), 194-202. https://doi.org/10.1016/j.tins.2007.03.006

Blanco-Pozo, M., Akam, T., & Walton, M. (2021). Dopamine reports reward prediction errors, but does not update policy, during inference-guided choice. *biorRxiv*. https://doi.org/10.1101/2021.06.25.449995

Bornstein, A. M., & Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current opinion in neurobiology*, *21*(3), 374-380. https://doi.org/10.1016/j.conb.2011.02.009

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262-280. https://doi.org/10.1016/j.cognition.2008.08.011

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263-1268. https://doi.org/10.1038/nn1525

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, *68*(5), 815-834. https://doi.org/10.1016/j.neuron.2010.11.022

Brown, H. D., McCutcheon, J. E., Cone, J. J., Ragozzino, M. E., & Roitman, M. F. (2011). Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. *Eur J Neurosci*, *34*(12), 1997-2006. https://doi.org/10.1111/j.1460-9568.2011.07914.x

Brozoski, T. J., Brown, R. M., Rosvold, H. E., & Goldman, P. S. (1979). Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey. *Science*, *205*(4409), 929-932. https://doi.org/10.1126/science.112679

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *The Psychological Review*, *58*(5), 313-323.

Busse, L., Ayaz, A., Dhruv, N. T., Katzner, S., Saleem, A. B., Scholvinck, M. L., Zaharia, A. D., & Carandini, M. (2011). The detection of visual contrast in the behaving mouse. *Journal of Neuroscience*, *31*(31), 11351-11361. https://doi.org/10.1523/JNEUROSCI.6689-10.2011

Carlsson, A., Lindquist, M., & Magnusson, T. (1957). 3,4-Dihydroxyphenylalanine and 5-hydroxytryptophan as reserpine antagonists. *Nature*, *180*, 1200.

Carr, D. B., & Sesack, S. R. (2000). Projections from the rat prefrontal cortex to the ventral tegmental area: target specificity in the synaptic associations with mesoaccumbens and mesocortical neurons. *Journal of Neuroscience*, *20*(10), 3864-3873. https://doi.org/10.1523/JNEUROSCI.20-10-03864.2000

Chen, T. W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., & Kim, D. S. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, *499*(7458), 295-300. https://doi.org/10.1038/nature12354

Coddington, L. T., & Dudman, J. T. (2018). The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nature Neuroscience*, *21*(11), 1563-1573. https://doi.org/10.1038/s41593-018-0245-7

Coddington, L. T., & Dudman, J. T. (2019). Learning from action: reconsidering movement signaling in midbrain dopamine neuron activity. *Neuron*, *104*(1), 63-77. https://doi.org/10.1016/j.neuron.2019.08.036

Coddington, L. T., Lindo, S. E., & Dudman, J. T. (2021). Mesolimbic dopamine adapts the rate of learning from errors in performance. *biorRxiv*. https://doi.org/10.1101/2021.05.31.446464

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, *482*(7383), 85-88. https://doi.org/10.1038/nature10754

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576-586. https://doi.org/10.1038/s41583-020-0355-6

Corbett, D., & Wise, R. A. (1980). Intracranial self-stimulation in relation to the ascending dopaminergic systems of the midbrain: a moveable electrode mapping study. *Brain research*, *185*(1), 1-15. https://doi.org/10.1016/0006-8993(80)90666-6

Cotzias, G. C., Van Woert, M. H., & Schiffer, L. M. (1967). Aromatic amino acids and modification of parkinsonism. *New England Journal of Medicine*, *276*(7), 374-379.

Cox, J., & Witten, I. B. (2019). Striatal circuits for reward learning and decision-making. *Nature Reviews Neuroscience*, *20*(8), 482-494. https://doi.org/10.1038/s41583-019-0189-2

da Silva, J. A., Tecuapetla, F., Paixao, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, *554*(7691), 244-248. https://doi.org/10.1038/nature25457

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671-675. https://doi.org/10.1038/s41586-019-1924-6

Darvas, M., Wunsch, A. M., Gibbs, J. T., & Palmiter, R. D. (2014). Dopamine dependency for acquisition and performance of Pavlovian conditioned response. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(7), 2764-2769. https://doi.org/10.1073/pnas.1400332111

Davis, K. L., Kahn, R. S., Ko, G., & Davidson, M. (1991). Dopamine in schizophrenia: a review and reconceptualization. *American Journal of Psychiatry*, *148*(11), 1474-1486. https://doi.org/10.1176/ajp.148.11.1474

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704-1711. https://doi.org/10.1038/nn1560

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185-196. https://doi.org/10.1016/j.conb.2008.08.003

de Jong, J. W., Afjei, S. A., Pollak Dorocic, I., Peck, J. R., Liu, C., Kim, C. K., Tian, L., Deisseroth, K., & Lammel, S. (2019). A neural circuit mechanism for encoding aversive stimuli in the mesolimbic dopamine system. *Neuron*, *101*(1), 133-151 e137. https://doi.org/10.1016/j.neuron.2018.11.005

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848-858. https://doi.org/10.1177/0956797616639301

Dickinson, A., & Balleine, B. W. (1994). Motivational control of goal-directed action. *Animal learning & behavior*, *22*, 1-18. https://doi.org/10.3758/BF03199951

Dodson, P. D., Dreyer, J. K., Jennings, K. A., Syed, E. C., Wade-Martins, R., Cragg, S. J., Bolam, J. P., & Magill, P. J. (2016). Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15), E2180-2188. https://doi.org/10.1073/pnas.1515941113

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology*, *30*(15), R860-R865. https://doi.org/10.1016/j.cub.2020.06.051

Durstewitz, D., Kroner, S., & Gunturkun, O. (1999). The dopaminergic innervation of the avian telencephalon. *Progress in Neurobiology*, *59*(2), 161-195. https://doi.org/10.1016/s0301-0082(98)00100-2

Egerton, A., Mehta, M. A., Montgomery, A. J., Lappin, J. M., Howes, O. D., Reeves, S. J., Cunningham, V. J., & Grasby, P. M. (2009). The dopaminergic basis of human behaviors: a review of molecular imaging studies. *Neuroscience & Biobehavioral Reviews*, *33*(7), 1109-1132. https://doi.org/10.1016/j.neubiorev.2009.05.005

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., Koay, S. A., Thiberge, S. Y., Daw, N. D., Tank, D. W., & Witten, I. B. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature*, *570*(7762), 509-513. https://doi.org/10.1038/s41586-019-1261-9

Farassat, N., Costa, K. M., Stojanovic, S., Albert, S., Kovacheva, L., Shin, J., Egger, R., Somayaji, M., Duvarci, S., Schneider, G., & Roeper, J. (2019). In vivo functional diversity of midbrain dopamine neurons within identified axonal projections. *eLife*, *8*. https://doi.org/10.7554/eLife.48408

Farrell, K., Lak, A., & Saleem, A. B. (2021). Midbrain dopamine neurons provide teaching signals for goal-directed navigation. *biorRxiv*. https://doi.org/10.1101/2021.02.17.431585

Faure, A., Haberland, U., Conde, F., & El Massioui, N. (2005). Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *Journal of Neuroscience*, *25*(11), 2771-2780. https://doi.org/10.1523/JNEUROSCI.3894-04.2005

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898-1902. https://doi.org/10.1126/science.1077349

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E., & Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. *Nature*, *469*(7328), 53-57. https://doi.org/10.1038/nature09588

Floresco, S. B., Magyar, O., Ghods-Sharifi, S., Vexelman, C., & Tse, M. T. (2006). Multiple dopamine receptor subtypes in the medial prefrontal cortex of the rat regulate set-shifting. *Neuropsychopharmacology*, *31*(2), 297-309. https://doi.org/10.1038/sj.npp.1300825

Gershman, S. J. (2014). Dopamine ramps are a consequence of reward prediction errors. *Neural computation*, *26*(3), 467-471. https://doi.org/10.1162/NECO_a_00559

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 3*(Supplement_3), 15647-15654. https://doi.org/10.1073/pnas.1014269108

Grace, A. A., & Bunney, B. S. (1984). The control of firing pattern in nigral dopamine neurons: burst firing. *Journal of Neuroscience*, *4*(11), 2877-2890. https://doi.org/10.1523/JNEUROSCI.04-11-02877.1984

Grospe, G. M., Baker, P. M., & Ragozzino, M. E. (2018). Cognitive flexibility deficits following 6-OHDA lesions of the rat dorsomedial striatum. *Neuroscience*, *374*, 80-90. https://doi.org/10.1016/j.neuroscience.2018.01.032

Grothe, B., & Pecka, M. (2014). The natural history of sound localization in mammals–a story of neuronal inhibition. *Frontiers in neural circuits*, *8*(65), 116. https://doi.org/10.3389/fncir.2014.00116

Guru, A., Seo, C., Post, R. J., Kullakanda, D. S., Schaffer, J. A., & Warden, M. R. (2020). Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map [Word Document]. *bioRxiv*, *542*, aau8722-8740. https://doi.org/10.1101/2020.05.21.108886

Hamid, A. A., Frank, M. J., & Moore, C. I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell*, *184*(10), 2733-2749 e2716. https://doi.org/10.1016/j.cell.2021.03.046

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., Kennedy, R. T., Aragona, B. J., & Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, *19*(1), 117-126. https://doi.org/10.1038/nn.4173

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Rio, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J Neurosci*, *34*(3), 698-704. https://doi.org/10.1523/JNEUROSCI.2489-13.2014

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245-258. https://doi.org/10.1016/j.neuron.2017.06.011

Hayar, A., Bryant, J. L., Boughter, J. D., & Heck, D. H. (2006). A low-cost solution to measure mouse licking in an electrophysiological setup with a standard analog-to-digital converter. *Journal of Neuroscience Methods*, *153*(2), 203-207. https://doi.org/10.1016/j.jneumeth.2005.10.023

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304-309. https://doi.org/10.1038/1124

Howe, M. W., & Dombeck, D. A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature*, *535*(7613), 505-510. https://doi.org/10.1038/nature18942

Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E., & Graybiel, A. M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, *500*(7464), 575-579. https://doi.org/10.1038/nature12475

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: version III–the final common pathway. *Schizophrenia Bulletin*, *35*(3), 549-562. https://doi.org/10.1093/schbul/sbp006

Hughes, R. N., Bakhurin, K. I., Petter, E. A., Watson, G. D. R., Kim, N., Friedman, A. D., & Yin, H. H. (2020). Ventral tegmental dopamine neurons control the impulse vector during motivated behavior. *Current Biology*, *30*(14), 2681-2694 e2685. https://doi.org/10.1016/j.cub.2020.05.003

Hull, C. L. (1943). *Principles of behavior: an introduction to behavior theory*. Appleton-Century.

Hunnicutt, B. J., Jongbloets, B. C., Birdsong, W. T., Gertz, K. J., Zhong, H., & Mao, T. (2016). A comprehensive excitatory input map of the striatum reveals novel functional organization. *eLife*, *5*. https://doi.org/10.7554/eLife.19103

Hwang, J., Mitz, A. R., & Murray, E. A. (2019). NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB. *Journal of Neuroscience Methods*, *323*, 13-21. https://doi.org/10.1016/j.jneumeth.2019.05.002

Hyland, B. I., Reynolds, J. N., Hay, J., Perk, C. G., & Miller, R. (2002). Firing modes of midbrain dopamine cells in the freely moving rat. *Neuroscience*, *114*(2), 475-492. https://doi.org/10.1016/s0306-4522(02)00267-1

Ito, M., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current opinion in neurobiology*, *21*(3), 368-373. https://doi.org/10.1016/j.conb.2011.04.001

Jacob, S. N., Ott, T., & Nieder, A. (2013). Dopamine regulates two classes of primate prefrontal neurons that represent sensory signals. *Journal of Neuroscience*, *33*(34), 13724-13734. https://doi.org/10.1523/JNEUROSCI.0210-13.2013

Jas, M., Achakulvisut, T., Idrizović, A., Acuna, D., Antalek, M., Marques, V., Odland, T., Garg, R., Agrawal, M., Umegaki, Y., Foley, P., Fernandes, H., Harris, D., Li, B., Pieters, O., Otterson, S., De Toni, G., Rodgers, C., Dyer, E., . . . Ramkumar, P. (2020). Pyglmnet: Python implementation of elastic-net regularized generalized linear models. *Journal of Open Source Software*, *5*(47), 1959-1953. https://doi.org/10.21105/joss.01959

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*(4-6), 535-547. https://doi.org/10.1016/s0893-6080(02)00047-3

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Keiflin, R., Pribut, H. J., Shah, N. B., & Janak, P. H. (2019). Ventral tegmental dopamine neurons participate in reward identity predictions. *Current Biology*, *29*(1), 93-103 e103. https://doi.org/10.1016/j.cub.2018.11.050

Kim, H. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., & Gershman, S. J. (2020). A unified framework for dopamine signals across timescales. *Cell*, *183*(6), 1600-1616. e1625. https://doi.org/10.1016/j.cell.2020.11.013

Kosillo, P., Zhang, Y. F., Threlfell, S., & Cragg, S. J. (2016). Cortical control of striatal dopamine transmission via striatal cholinergic interneurons. *Cerebral Cortex*, *26*(11), 4160-4169. https://doi.org/10.1093/cercor/bhw252

Kutlu, M. G., Zachry, J. E., Melugin, P. R., Cajigas, S. A., Chevee, M. F., Kelley, S. J., Kutlu, B., Tian, L., Siciliano, C. A., & Calipari, E. S. (2021). Dopamine release in the nucleus accumbens core signals perceived saliency. *Current Biology*. https://doi.org/10.1016/j.cub.2021.08.052

Lak, A., Hueske, E., Hirokawa, J., Masset, P., Ott, T., Urai, A. E., Donner, T. H., Carandini, M., Tonegawa, S., Uchida, N., & Kepecs, A. (2020). Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *eLife*, *9*. https://doi.org/10.7554/eLife.49834

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, *27*(6), 821-832. https://doi.org/10.1016/j.cub.2017.02.026

Lak, A., Okun, M., Moss, M. M., Gurnani, H., Farrell, K., Wells, M. J., Reddy, C. B., Kepecs, A., Harris, K. D., & Carandini, M. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron*, *105*(4), 700-711 e706. https://doi.org/10.1016/j.neuron.2019.11.018

Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2343-2348. https://doi.org/10.1073/pnas.1321596111

Lak, A., Stauffer, W. R., & Schultz, W. (2016). Dopamine neurons learn relative chosen value from probabilistic rewards. *eLife*, *5*. https://doi.org/10.7554/eLife.18044

Lammel, S., Ion, D. I., Roeper, J., & Malenka, R. C. (2011). Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron*, *70*(5), 855-862. https://doi.org/10.1016/j.neuron.2011.03.025

Lammel, S., Steinberg, E. E., Foldy, C., Wall, N. R., Beier, K., Luo, L., & Malenka, R. C. (2015). Diversity of transgenic mouse models for selective targeting of midbrain dopamine neurons. *Neuron*, *85*(2), 429-438. https://doi.org/10.1016/j.neuron.2014.12.036

Laubach, M., Amarante, L. M., Swanson, K., & White, S. R. (2018). What, if anything, is rodent prefrontal cortex? *eneuro*, *5*(5). https://doi.org/10.1523/ENEURO.0315-18.2018

Lee, D., Creed, M., Jung, K., Stefanelli, T., Wendler, D. J., Oh, W. C., Mignocchi, N. L., Luscher, C., & Kwon, H. B. (2017). Temporally precise labeling and control of neuromodulatory circuits in the mammalian brain. *Nature Methods*, *14*(5), 495-503. https://doi.org/10.1038/nmeth.4234

Lee, R. S., Mattar, M. G., Parker, N. F., Witten, I. B., & Daw, N. D. (2019). Reward prediction error does not explain movement selectivity in DMS-projecting dopamine neurons. *eLife*, *8*. https://doi.org/10.7554/eLife.42992

Lerner, T. N., Holloway, A. L., & Seiler, J. L. (2020). Dopamine, updated: reward prediction error and beyond. *Current opinion in neurobiology*, *67*, 123-130. https://doi.org/10.1016/j.conb.2020.10.012

Lerner, T. N., Shilyansky, C., Davidson, T. J., Evans, K. E., Beier, K. T., Zalocusky, K. A., Crow, A. K., Malenka, R. C., Luo, L., Tomer, R., & Deisseroth, K. (2015). Intact-brain analyses reveal distinct information carried by SNc dopamine subcircuits. *Cell*, *162*(3), 635-647. https://doi.org/10.1016/j.cell.2015.07.014

Liu, H., Zakiniaeiz, Y., Cosgrove, K. P., & Morris, E. D. (2019). Toward whole-brain dopamine movies: a critical review of PET imaging of dopamine transmission in the striatum and cortex. *Brain Imaging and Behavior*, *13*(2), 314-322. https://doi.org/10.1007/s11682-017-9779-7

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*(1), 145-163. https://doi.org/10.1152/jn.1992.67.1.145

Maes, E. J. P., Sharpe, M. J., Usypchuk, A. A., Lozzi, M., Chang, C. Y., Gardner, M. P. H., Schoenbaum, G., & Iordanova, M. D. (2020). Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors. *Nature Neuroscience*, *23*(2), 176-178. https://doi.org/10.1038/s41593-019-0574-1

Margolis, E. B., Lock, H., Hjelmstad, G. O., & Fields, H. L. (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *Journal of Physiology*, *577*(Pt 3), 907-924. https://doi.org/10.1113/jphysiol.2006.117069

Marshall, J. F., Levitan, D., & Stricker, E. M. (1976). Activation-induced restoration of sensorimotor functions in rats with dopamine-depleting brain lesions. *Journal of comparative and physiological psychology*, *90*(6), 536. https://doi.org/10.1037/h0077230

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281-1289. https://doi.org/10.1038/s41593-018-0209-y

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(7248), 837-841. https://doi.org/10.1038/nature08028

Matsumoto, M., & Takada, M. (2013). Distinct representations of cognitive and motivational signals in midbrain dopamine neurons. *Neuron*, *79*(5), 1011-1024. https://doi.org/10.1016/j.neuron.2013.07.002

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, *445*.

Menegas, W., Akiti, K., Amo, R., Uchida, N., & Watabe-Uchida, M. (2018). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature Neuroscience*, *21*(10), 1421-1430. https://doi.org/10.1038/s41593-018-0222-1

Menegas, W., Babayan, B. M., Uchida, N., & Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife*, *6*. https://doi.org/10.7554/eLife.21886

Menegas, W., Bergan, J. F., Ogawa, S. K., Isogai, Y., Umadevi Venkataraju, K., Osten, P., Uchida, N., & Watabe-Uchida, M. (2015). Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *eLife*, *4*, e10032. https://doi.org/10.7554/eLife.10032

Millan, M. J., Agid, Y., Brune, M., Bullmore, E. T., Carter, C. S., Clayton, N. S., Connor, R., Davis, S., Deakin, B., DeRubeis, R. J., Dubois, B., Geyer, M. A., Goodwin, G. M., Gorwood, P., Jay, T. M., Joels, M., Mansuy, I. M., Meyer-Lindenberg, A., Murphy, D., . . . Young, L. J. (2012). Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. *Nature Reviews Drug Discovery*, *11*(2), 141-168. https://doi.org/10.1038/nrd3628

Miller, K. J., Ludvig, E. A., Pezzulo, G., & Shenhav, A. (2018). Re-aligning models of habitual and goal-directed decision-making. In *Goal-directed decision making* (pp. 407-428). Academic Press.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, *72*(2), 1024-1027. https://doi.org/10.1152/jn.1994.72.2.1024

Mirenowicz, J., & Schultz, W. (1996). Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, *379*(6564), 449-451. https://doi.org/10.1038/379449a0

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533. https://doi.org/10.1038/nature14236

Mohebi, A., Pettibone, J. R., Hamid, A. A., Wong, J. T., Vinson, L. T., Patriarchi, T., Tian, L., Kennedy, R. T., & Berke, J. D. (2019). Dissociable dopamine dynamics for learning and motivation. *Nature*, *570*(7759), 65-70. https://doi.org/10.1038/s41586-019-1235-y

Mongia, S., Yamaguchi, T., Liu, B., Zhang, S., Wang, H., & Morales, M. (2019). The ventral tegmental area has calbindin neurons with the capability to co-release glutamate and dopamine into the nucleus accumbens. *European Journal of Neuroscience*, *50*(12), 3968-3984. https://doi.org/10.1111/ejn.14493

Montagu, K. A. (1957). Catechol compounds in rat tissues and in brains of different animals. *Nature*, *180*(4579), 244-245. https://doi.org/10.1038/180244a0

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, *9*(8), 1057-1063. https://doi.org/10.1038/nn1743

Muller, A., Joseph, V., Slesinger, P. A., & Kleinfeld, D. (2014). Cell-based reporters reveal in vivo dynamics of dopamine and norepinephrine release in murine cortex. *Nature Methods*, *11*(12), 1245-1252. https://doi.org/10.1038/nmeth.3151

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677-1686. https://doi.org/10.1038/s41593-019-0502-4

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, *14*(7), 2152-2176. https://doi.org/10.1038/s41596-019-0176-0

Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, *1*(3), 133-143. https://doi.org/10.1038/s42256-019-0025-4

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507-520. https://doi.org/10.1007/s00213-006-0502-4

Noudoost, B., & Moore, T. (2011). The role of neuromodulators in selective attention. *Trends in Cognitive Sciences*, *15*(12), 585-591. https://doi.org/10.1016/j.tics.2011.10.006

Nutt, D. J., Lingford-Hughes, A., Erritzoe, D., & Stokes, P. R. (2015). The dopamine theory of addiction: 40 years of highs and lows. *Nature Reviews Neuroscience*, *16*(5), 305-312. https://doi.org/10.1038/nrn3939

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452-454. https://doi.org/10.1126/science.1094285

Olds, J., Killam, K. F., & Bach-Y-Rita, P. (1956). Self-stimulation of the brain used as a screening method for tranquilizing drugs. *Science*, *124*(3215), 265-266. https://doi.org/10.1126/science.124.3215.265

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, *47*, 419-427.

Olds, M. E., & Olds, J. (1969). Effects of lesions in medial forebrain bundle on self-stimulation behavior. *American Journal of Physiology*, *217*(5), 1253-1264. https://doi.org/10.1152/ajplegacy.1969.217.5.1253

Ott, T., Jacob, S. N., & Nieder, A. (2014). Dopamine receptors differentially enhance rule coding in primate prefrontal cortex neurons. *Neuron*, *84*(6), 1317-1328. https://doi.org/10.1016/j.neuron.2014.11.012

Pan, W. X., Coddington, L. T., & Dudman, J. T. (2021). Dissociable contributions of phasic dopamine activity to reward and prediction. *Cell Reports*, *36*(10), 109684. https://doi.org/10.1016/j.celrep.2021.109684

Pan, W. X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *Journal of Neuroscience*, *25*(26), 6235-6242. https://doi.org/10.1523/JNEUROSCI.1478-05.2005

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., Daw, N. D., & Witten, I. B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience*, *19*(6), 845-854. https://doi.org/10.1038/nn.4287

Patriarchi, T., Cho, J. R., Merten, K., Howe, M. W., Marley, A., Xiong, W. H., Folk, R. W., Broussard, G. J., Liang, R., Jang, M. J., Zhong, H., Dombeck, D., von Zastrow, M., Nimmerjahn, A., Gradinaru, V., Williams, J. T., & Tian, L. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science*, *360*(6396), eaat4422. https://doi.org/10.1126/science.aat4422

Patriarchi, T., Mohebi, A., Sun, J., Marley, A., Liang, R., Dong, C., Puhger, K., Mizuno, G. O., Davis, C. M., Wiltgen, B., von Zastrow, M., Berke, J. D., & Tian, L. (2020). An expanded palette of dopamine sensors for multiplex imaging in vivo. *Nature Methods*, *17*(11), 1147-1155. https://doi.org/10.1038/s41592-020-0936-3

Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Paxinos, G., & Franklin, K. B. J. (2001). *The mouse brain in stereotaxic coordinates* (2nd Edition ed.). Academic Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825--2830.

Phillips, M. I., & Olds, J. (1969). Unit activity: motiviation-dependent responses from midbrain neurons. *Science*, *165*(3899), 1269-1271. https://doi.org/10.1126/science.165.3899.1269

Poulin, J. F., Caronia, G., Hofer, C., Cui, Q., Helm, B., Ramakrishnan, C., Chan, C. S., Dombeck, D. A., Deisseroth, K., & Awatramani, R. (2018). Mapping projections of molecularly defined dopamine neuron subtypes using intersectional genetic approaches. *Nature Neuroscience*, *21*(9), 1260-1271. https://doi.org/10.1038/s41593-018-0203-4

Puig, M. V., & Miller, E. K. (2012). The role of prefrontal dopamine D1 receptors in the neural mechanisms of associative learning. *Neuron*, *74*(5), 874-886. https://doi.org/10.1016/j.neuron.2012.04.018

Ranganath, A., & Jacob, S. N. (2016). Doping the mind: dopaminergic modulation of prefrontal cortical cognition. *Neuroscientist*, *22*(6), 593-603. https://doi.org/10.1177/1073858415602850

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M. R., & Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews Neuroscience*, *11*(11), 760-772. https://doi.org/10.1038/nrn2915

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*, 151-182.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton Century Crofts.

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67-70. https://doi.org/10.1038/35092560

Robinson, D. L., Venton, B. J., Heien, M. L., & Wightman, R. M. (2003). Detecting subsecond dopamine release with fast-scan cyclic voltammetry in vivo. *Clinical Chemistry*, *49*(10), 1763-1773. https://doi.org/10.1373/49.10.1763

Romo, R., & Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, *63*(3), 592-606. https://doi.org/10.1152/jn.1990.63.3.592

Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D., & Pillow, J. W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron*, *109*(4), 597-610 e596. https://doi.org/10.1016/j.neuron.2020.12.004

Russo, S. J., & Nestler, E. J. (2013). The brain reward circuitry in mood disorders. *Nature Reviews Neuroscience*, *14*(9), 609-625. https://doi.org/10.1038/nrn3381

Salamone, J. D., & Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. *Neuron*, *76*(3), 470-485. https://doi.org/10.1016/j.neuron.2012.10.021

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*(6537). https://doi.org/10.1126/science.abf4740

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat Rev Neurosci*, *10*(12), 885-892. https://doi.org/10.1038/nrn2753

Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *Journal of Neurophysiology*, *56*(5), 1439-1461. https://doi.org/10.1152/jn.1986.56.5.1439

Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual Review of Neuroscience*, *30*, 259-288. https://doi.org/10.1146/annurev.neuro.28.061604.135722

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*(3), 900-913. https://doi.org/10.1523/JNEUROSCI.13-03-00900.1993

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., & Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, *63*(3), 607-624. https://doi.org/10.1152/jn.1990.63.3.607

Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*.

Sharpe, M. J., Batchelor, H. M., Mueller, L. E., Yun Chang, C., Maes, E. J. P., Niv, Y., & Schoenbaum, G. (2020). Dopamine transients do not act as model-free prediction errors during associative learning. *Nature Communications*, *11*(1), 106. https://doi.org/10.1038/s41467-019-13953-1

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354-359. https://doi.org/10.1038/nature24270

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis*. Appleton-Century.

Skinner, B. F. (1953). *Science and human behavior*. Macmillan.

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, *20*(4), 581-589. https://doi.org/10.1038/nn.4520

Stauffer, W. R., Lak, A., Yang, A., Borel, M., Paulsen, O., Boyden, E. S., & Schultz, W. (2016). Dopamine neuron-specific optogenetic stimulation in rhesus macaques. *Cell*, *166*(6), 1564-1571 e1566. https://doi.org/10.1016/j.cell.2016.08.024

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966-973. https://doi.org/10.1038/nn.3413

Sun, F., Zeng, J., Jing, M., Zhou, J., Feng, J., Owen, S. F., Luo, Y., Li, F., Wang, H., Yamaguchi, T., Yong, Z., Gao, Y., Peng, W., Wang, L., Zhang, S., Du, J., Lin, D., Xu, M., Kreitzer, A. C., . . . Li, Y. (2018). A genetically encoded fluorescent sensor enables rapid and specific detection of dopamine in flies, fish, and mice. *Cell*, *174*(2), 481-496 e419. https://doi.org/10.1016/j.cell.2018.06.042

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9-44. https://doi.org/10.1007/bf00115009

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(2), 135-170. https://doi.org/10.1037/0033-295X.88.2.135

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction* (second ed.). MIT Press.

Swanson, J. M., Kinsbourne, M., Nigg, J., Lanphear, B., Stefanatos, G. A., Volkow, N., Taylor, E., Casey, B. J., Castellanos, F. X., & Wadhwa, P. D. (2007). Etiologic subtypes of attention-deficit/hyperactivity disorder: brain imaging, molecular genetic and environmental factors and the dopamine hypothesis. *Neuropsychology Review*, *17*(1), 39-59. https://doi.org/10.1007/s11065-007-9019-9

Syed, E. C., Grima, L. L., Magill, P. J., Bogacz, R., Brown, P., & Walton, M. E. (2016). Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nature Neuroscience*, *19*(1), 34-36. https://doi.org/10.1038/nn.4187

Takahashi, Y. K., Roesch, M. R., Wilson, R. C., 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*(12), 1590-1597. https://doi.org/10.1038/nn.2957

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human pavlovian-instrumental transfer. *Journal of Neuroscience*, *28*(2), 360-368. https://doi.org/10.1523/JNEUROSCI.4028-07.2008

Tan, K. R., Yvon, C., Turiault, M., Mirzabekov, J. J., Doehner, J., Labouebe, G., Deisseroth, K., Tye, K. M., & Lüscher, C. (2012). GABA neurons of the VTA drive conditioned place aversion. *Neuron*, *73*(6), 1173-1183. https://doi.org/10.1016/j.neuron.2012.02.015

Thorn, C. A., Atallah, H., Howe, M., & Graybiel, A. M. (2010). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, *66*(5), 781-795. https://doi.org/10.1016/j.neuron.2010.04.036

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review*, *2*(4).

Threlfell, S., Lalic, T., Platt, N. J., Jennings, K. A., Deisseroth, K., & Cragg, S. J. (2012). Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. *Neuron*, *75*(1), 58-64. https://doi.org/10.1016/j.neuron.2012.04.038

Tidey, J. W., & Miczek, K. A. (1996). Social defeat stress selectively alters mesocorticolimbic dopamine release: an in vivo microdialysis study. *Brain research*, *721*(1-2), 140-149. https://doi.org/10.1016/0006-8993(96)00159-x

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642-1645. https://doi.org/10.1126/science.1105370

Tolman, E. C. (1948). Cognitive maps in rats and men. *The Psychological Review*, *55*(4), 189-208. https://doi.org/10.1037/h0061626

Treadway, M. T., & Zald, D. H. (2011). Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neuroscience & Biobehavioral Reviews*, *35*(3), 537-555. https://doi.org/10.1016/j.neubiorev.2010.06.006

Tsai, H. C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, *324*(5930), 1080-1084. https://doi.org/10.1126/science.1168878

Tsutsui-Kimura, I., Matsumoto, H., Akiti, K., Yamada, M. M., Uchida, N., & Watabe-Uchida, M. (2020). Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. *eLife*, *9*. https://doi.org/10.7554/eLife.62390

Ungerstedt, U. (1971). Adipsia and aphagia after 6-hydroxydopamine induced degeneration of the nigro-striatal dopamine system. *Acta Physiologica Scandinavica*, *82*(S367), 95-122. https://doi.org/10.1111/j.1365-201x.1971.tb11001.x

Van Rossum, J. (1966). The significance of dopamine-receptor blockade for the mechanism of action of neuroleptic drugs. *Archives internationales de pharmacodynamie et de therapie*, *160*(2), 492-494.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy, C. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2

Volkow, N. D., & Morales, M. (2015). The brain on drugs: from reward to addiction. *Cell*, *162*(4), 712-725. https://doi.org/10.1016/j.cell.2015.07.046

Volkow, N. D., Wang, G. J., Kollins, S. H., Wigal, T. L., Newcorn, J. H., Telang, F., Fowler, J. S., Zhu, W., Logan, J., Ma, Y., Pradhan, K., Wong, C., & Swanson, J. M. (2009). Evaluating dopamine reward pathway in ADHD: clinical implications. *JAMA*, *302*(10), 1084-1091. https://doi.org/10.1001/jama.2009.1308

Wang, D. V., Viereckel, T., Zell, V., Konradsson-Geuken, A., Broker, C. J., Talishinsky, A., Yoo, J. H., Galinato, M. H., Arvidsson, E., Kesner, A. J., Hnasko, T. S., Wallen-Mackenzie, A., & Ikemoto, S. (2017). Disrupting glutamate co-transmission does not affect acquisition of conditioned behavior reinforced by dopamine neuron activation. *Cell Reports*, *18*(11), 2584-2591. https://doi.org/10.1016/j.celrep.2017.02.062

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural circuitry of reward prediction error. *Annual Review of Neuroscience*, *40*, 373-394. https://doi.org/10.1146/annurev-neuro-072116-031109

Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A., & Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron*, *74*(5), 858-873. https://doi.org/10.1016/j.neuron.2012.03.017

Wei, W., Mohebi, A., & Berke, J. (2021). Striatal dopamine pulses follow a temporal discounting spectrum. *biorRxiv*. https://doi.org/10.1101/2021.10.31.466705

Weinstein, J. J., Chohan, M. O., Slifstein, M., Kegeles, L. S., Moore, H., & Abi-Dargham, A. (2017). Pathway-specific dopamine abnormalities in schizophrenia. *Biological Psychiatry*, *81*(1), 31-42. https://doi.org/10.1016/j.biopsych.2016.03.2104

Willuhn, I., Burgeno, L. M., Everitt, B. J., & Phillips, P. E. (2012). Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(50), 20703-20708. https://doi.org/10.1073/pnas.1213460109

Wise, R. A. (1985). The anhedonia hypothesis: mark III. *Behavioral and Brain Sciences*, *8*(1), 178-186. https://doi.org/10.1017/S0140525X00020306

Wise, R. A. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience*, *5*(6), 483-494. https://doi.org/10.1038/nrn1406

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(6204), 1616-1620. https://doi.org/10.1126/science.1255514

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464-476. https://doi.org/10.1038/nrn1919

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*(1), 181-189. https://doi.org/10.1111/j.1460-9568.2004.03095.x

Yin, H. H., Mulcare, S. P., Hilario, M. R., Clouse, E., Holloway, T., Davis, M. I., Hansson, A. C., Lovinger, D. M., & Costa, R. M. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nature Neuroscience*, *12*(3), 333-341. https://doi.org/10.1038/nn.2261

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513-523. https://doi.org/10.1111/j.1460-9568.2005.04218.x

Zell, V., Steinkellner, T., Hollon, N. G., Warlow, S. M., Souter, E., Faget, L., Hunker, A. C., Jin, X., Zweifel, L. S., & Hnasko, T. S. (2020). VTA glutamate neuron activity drives positive reinforcement absent dopamine co-release. *Neuron*, *107*(5), 864-873 e864. https://doi.org/10.1016/j.neuron.2020.06.011

Zolin, A., Cohn, R., Pang, R., Siliciano, A. F., Fairhall, A. L., & Ruta, V. (2021). Context-dependent representations of movement in Drosophila dopaminergic reinforcement pathways. *Nature Neuroscience*, *24*(11), 1555-1566. https://doi.org/10.1038/s41593-021-00929-y

8. Acknowledgements

The work presented in this dissertation would not have been possible without the help and support of several people that I would like to thank here. I would like to thank my supervisor Simon Jacob for providing a supportive lab environment to perform cutting-edge research, for his enthusiasm and ambition that motivated me to persist even in the face of challenges, for his scientific rigor and attention to detail that helped sharpen my thinking, and for his transdisciplinary interests and openness to ideas that encouraged me to pursue the unconventional side project of obtaining a master's degree in psychology during the time of my PhD. I would like to thank my thesis advisory committee members Anton Sirota and Mark Hübener for providing useful advice and guidance and for broadening the perspective on my research projects as they evolved over the years. I would like to thank all fellow Jacob lab members for their support and inspiration along my scientific journey. A special thanks goes to Ajit and Daniel for setting up our first lab and to Ajit, Daniel, and Leonie for setting up our second lab together with me, to Leonie for pioneering mouse behavior together with me, and to Beatrice for helping me with data collection and for continuing the dopamine project in the future!

I would like to thank the Graduate School of Systemic Neurosciences (GSN) for providing an interdisciplinary academic framework that enabled me to further my knowledge and skills while entering the field of neuroscience during the preparatory year of the fast-track PhD program. I would like to thank all staff members of the GSN for their kind support and flexibility as well as for their enthusiasm to shape the GSN not only as a stimulating research environment but also a fantastic professional and social network. I would like to thank my fellow GSN students for making the fast-track year a very enjoyable experience!

Finally, I would like to thank my family and friends for providing endless support both in times of excitement and enthusiasm as well as in times of challenges and doubts. I am eternally grateful to my parents and my brother for providing a loving environment to grow up in and for their unconditional support and encouragement that enabled all my academic achievements. Last but not least, I would like to thank the love of my life, Moni, for being there for me all the time, supporting me throughout ups and downs, and above all, for starting our own little family with me, which brings us so much joy!