EXPECTED UNCERTAINTY DRIVES PHASIC AND TONIC EXPLORATION IN RHESUS MACAQUES DURING PROBABILISTIC LEARNING

By

Sylvia Roth Bindas

THESIS

Presented to the Department of Behavioral Neuroscience
and the Oregon Health & Science University
School of Medicine

In partial fulfillment of the requirements for the degree of
Master of Science

June 1, 2021

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Sylvia Roth Bindas

"EXPECTED UNCERTAINTY DRIVES PHASIC AND TONIC EXPLORATION IN RHESUS MACAQUES DURING PROBABILISTIC LEARNING"

Has been approved by:

Matt Lattal, Ph.D.
Committee Chair

Vincent Costa, Ph.D. (Mentor)
Committee Member

Kathy Grant, Ph.D.
Committee Member

Marina Guizzetti, Ph.D.
Committee Member

Suzanne Mitchell, Ph.D.
Committee Member

Table of Contents

List of Figures

Acknowledgements

I would like to thank Dr. Vincent Costa for supervising my lab work at the Oregon National Primate Research Center. I would also like to thank Dr. Matt Lattal and Dr. Marina Guizzetti who served with Dr. Costa on my Thesis Advisory Committee and helped me prepare this thesis for submission. I would not have been able to complete this work without the original data, collected and reported on by Dr. Vincent Costa, Valery L. Tran, Dr. Janita Turchi, and Dr. Bruno Averbeck. Additionally, I am thankful for the financial support I received from the ARCS Foundation as well as the Ruth L. Kirschstein Predoctoral (T32) National Research Service Award.

I am deeply grateful for the support of my lab-mates, Dr. Jeremiah Morrow and Alexis Cooper. Their impossibly strong work ethic, composure, and compassion made our lab's world go round, and helped me to complete the work presented here.

I am also thankful to my siblings — Ava, Erica, and Evan — my parents — Jim and Nancy — and my perfect [cat] son – Étienne – for the overwhelming love and support they have given me over the past two years. Without it, this work would surely not be possible.

Abstract

In changing environments, adaptive decision-making requires balancing when to choose familiar, known options with when to explore new, unknown options. This balancing act, known as the explore-exploit tradeoff, is critical to how we make choices that can maximize reward. Specifically, exploration supports optimal decision-making by reducing the uncertainty associated with previously unknown choices. Exploration is often seen as phasic, where the decision to explore depends on peaks in uncertainty that signal when the benefit of exploring is greatest. However, exploration can also be tonic, occurring more regularly in time. While tonic exploration has been demonstrated in settings where uncertainty is limited to discrete, unexpected rule changes, it is unclear how tonic exploration relates to expected uncertainty from stochastic reward outcomes. Here, we use a Bayesian modeling approach to show that spontaneous errors (i.e. lapses) in the reversal phase of a two-armed bandit reversal learning task reflect a form of tonic exploration. This tonic exploration coexists with phasic exploration in the task and does not scale with environmental uncertainty. Further, we find that tonic exploration is directed rather than random, as lapses can be accurately predicted by Bayesian estimates of unpredictability and choice consistency. Our results demonstrate how tonic exploration complements phasic exploration in changing environments as a directed strategy to reduce uncertainty and maximize reward.

Introduction

To make adaptive choices in complex environments, decision-makers must weigh and evaluate what they know about their available options. One important consideration for decision-making is the *value* of a given option, or the mean expected outcome that will result from selecting it. Some options, which have been chosen frequently, have a rich history of outcome information that inform the estimated value of choosing that option (Sutton & Barto, 1998). For other options, which have not been chosen as frequently, a smaller number of outcomes results in a less certain estimate of the mean or expected value. This leads to a fundamental dilemma for decision-makers, known as the explore-exploit dilemma: is it better to choose a familiar option whose reward value is known, or to choose an unfamiliar option that is less known but potentially more valuable (Addicott et al., 2017; Averbeck, 2015)? To choose the familiar option – to exploit – maximizes rewards in the short-term, but it is relatively uninformative as nothing is learned about other potentially rewarding options. To choose the unfamiliar option – to explore – reduces uncertainty and can maximize reward in the long-term, but it may be risky and unsuccessful (Barack & Gold, 2016).

Understanding how humans balance exploration and exploitation also has implications for cases of maladaptive or pathological decision-making. For example, the compulsive habitual behavior seen in many substance use disorders (Everitt & Robbins, 2016) can be characterized in the framework of exploration and exploitation. In both alcohol use disorder and chronic cocaine use, substance use has been shown to lead to reduced exploratory behaviors as well as reduced feedback sensitivity (Morris et al., 2016; Zhukovsky et al., 2019), resulting in a tendency to over-exploit (i.e. perseverate) despite negative outcomes. In the field of psychiatry, while there is little evidence for a causal link between maladaptive exploration and psychiatric disease, individual tendencies to explore or exploit can serve as an additional dimensional construct for the classification of mental disorders (Addicott et al., 2017; Scholl & Klein-Flugge, 2018). For example, demonstrating that uncertainty-driven exploration is reduced in people diagnosed with schizophrenia (Strauss et al., 2011; Waltz et al., 2020) has provided new avenues for exploring the potential neural mechanisms that are altered in the disease. Thus, understanding the mechanisms that govern exploratory behavior may prove useful in treating addiction and in understanding psychiatric disease.

Despite a long history of research in many model systems (Krebs et al., 1978; Sims et al., 2008; Thatcher et al., 2019) and in fields from behavioral ecology to mathematics to psychiatry, there is no known optimal policy for trading off exploration and exploitation (Averbeck, 2015; Cohen et al., 2007; Gittins, 1979). From a computational perspective, this is because making an optimal decision (i.e. one that will maximize future reward in the long-term) requires a great deal of calculation. In theory, the best policy for exploring multiple options is to calculate the average value of *each* option for the current choice, the next choice, and so on over *all* possible futures even before the current choice is made (Feng et al., 2021). This calculation of future values, however, is incredibly taxing and requires a great deal of cognitive resources. Therefore, work examining the neural and behavioral bases of exploration focus on identifying simpler, more feasible strategies the brain implements instead of this optimal, but less tractable one (Averbeck, 2015; Daw et al., 2006, Sutton & Barto, 1998).

There are many decision-making paradigms used to probe the strategies that underlie explore-exploit decisions (Constantino and Daw, 2015; Costa et al., 2014; Glass et al., 2011; Wilson et al., 2014). In order to assess whether decisions are exploratory or exploitative, subjects must be presented with multiple options to choose from. Typically, these options have a predetermined and unknown (to the subject) reward structure which requires trial-by-trial learning from the subject to estimate the current and future reward values (Sutton & Barto, 1998). Importantly, the more a given option is chosen, the less uncertain subjects are about its current and future value (Wilson et al., 2014). When selecting between the multiple available options, then, there are two ways to approach maximizing long-term reward in the presence of uncertainty: subjects can use their current value estimates to choose the most rewarding option (and risk missing out on a better one) or they can further reduce uncertainty by choosing a less known option (but risk not being rewarded for it).

One very common example of a task used to study exploration is the n-armed bandit. Often compared to playing on multiple (*n*) adjacent slot machines, a multi-armed bandit task presents subjects with multiple possible actions that each have their own expected reward value (Sutton & Barto, 1998). Subjects in a bandit task are not told the values of choosing each option, rather they must form estimates by repeatedly sampling different options. Imagine, for example, a subject who is presented with a row of four slot machines and has only one hour to play them. In order to maximize reward during that hour, the subject must decide which machines to play, how many times to play each machine, and when to bail on a machine that seems unlucky. In these ways, the bandit task forces subjects to balance exploration and exploitation to get the most reward in the long term (Sutton & Barto, 1998). As with other paradigms probing explore-exploit strategies, the bandit task leverages uncertainty to drive exploration.

Uncertainty in these paradigms can be subdivided into two different types (Bland & Schaefer, 2012; Soltani & Izquierdo, 2019; Yu & Dayan, 2005). The first, expected uncertainty, is related to the variability of choice outcomes based on our expectations of what could occur (Cohen et al., 2007b; Fiorillo et al., 2003; Paulus et al., 2004; Polezzi et al., 2008; Volz et al., 2003). Expected uncertainty can be thought of as the known risk of not getting rewarded, reflecting some unavoidable but known variability in an environment (Soltani & Izquierdo, 2019; Yu and Dayan, 2005). Unexpected uncertainty, on the other hand, is related to the variability of the environment or rules themselves (Behrens et al., 2007; Courville et al., 2006; Doya, 2008; Krugel et al., 2009; Nassar et al., 2010; Payzan-LeNestour & Bossaerts, 2011; Piray & Daw, 2020a; Piray & Daw, 2020b; Rushworth & Behrens, 2008). Importantly, unexpected uncertainty reflects a violation of the subject's estimate of expected uncertainty in their environment (Soltani & Izquierdo, 2019; Yu and Dayan, 2005).

Both of these types of uncertainty are fundamental in tipping the balance between exploration and exploitation (Badre et al., 2012; Gershman, 2018; Schulz & Gershman, 2019). A powerful method for understanding how this happens is to generate mathematical models that quantify uncertainty and then see how well those models are able to describe behavior. For example, in pioneering work by Daw and colleagues (2006), experimenters proposed three separate mathematical models that might explain how subjects guide their choices to explore or exploit in an unstable environment. In the first model, known as an 'epsilon-greedy' model (Sutton and Barto, 1998), decision-makers keep track of the value of each option and usually exploit by choosing the one with the highest value, but at some fixed rate (specified by an 'epsilon' value) they explore and randomly select another option. In the second model, known as the 'softmax' model, decision-makers choose each option at a rate proportional to its expected value (also known as probability matching), and thus are more likely to choose the most rewarding option. However, this tendency to exploit is 'softened' by an additional parameter that accounts for different sensitivities to the contrast in values between options; when this parameter, often referred to as the temperature, is low, subjects' choices are highly exploitative and constrained by relative value. In contrast, when the temperature is high, decision-makers respond more randomly with respect to relative value, leading them to explore more. Lastly, the third model was identical to the second with the exception of one added parameter, a so-called ‘uncertainty-bonus’ that increases the probability of choosing options whose outcomes are less certain because they have not been selected as often.

After human participants performed a four-armed bandit task, Daw et al. (2006) compared their choice behavior to predictions from each of the three models and found that their behavior was best fit by the softmax decision model. By examining how well theoretical or mathematical solutions to the explore-exploit dilemma mapped onto observed human behavior, Daw et al. (2006) provided critical insight into how humans are solving this problem. Further, the mathematical models used provide concrete parameter estimates that can facilitate the identification of brain regions helping to solve the explore-exploit dilemma. For example, in the same work by Daw and colleagues (2006), they showed that activity in the ventromedial prefrontal cortex (vmPFC) scaled positively with the model’s estimated expected value for a given choice, where activity in the dorsolateral prefrontal cortex (dlPFC) scaled negatively with the same parameter. Overall, these findings demonstrate the utility of mathematical models in characterizing the computations and mechanisms that allow humans and animals to navigate the explore-exploit dilemma.

Most computational models attempting to describe a solution to the explore-exploit dilemma quantify uncertainty in one of two ways. First, some models introduce an explicit bias towards information that is often expressed as an added ‘bonus’ to the value of novel or uncertain options (Gershman, 2018). In this way, so-called directed exploration targets high-uncertainty options to gather information and ensure maximum reward in the long-term

(Gershman, 2018). The second class of models, known as random exploration algorithms, introduce random noise into choice behavior (Thompson, 1933; Wilson et al., 2014). Typically, this noise in behavior is set to scale with uncertainty in the environment. Therefore, where directed exploration is sensitive to the relative uncertainty of certain options, random exploration is sensitive to the total uncertainty of the environment (Gershman, 2018).

Humans have been shown to employ both random and directed exploration within the same task structure, as was first demonstrated by Wilson et al. (2014). In their study, Wilson and colleagues had participants play a series of games in which they made choices between two options with different probabilistic rewards. By accounting for the correlation between information (i.e. how much is known about an option based on how often it is sampled) and reward, investigators were able to parse both directed and random exploration. Where directed exploration was expressed as a bias towards information seeking over reward, random exploration was expressed as randomness in choice behavior (Wilson et al., 2014). More recently, computational modeling has supported that hybrids of random and directed algorithms for exploration most efficiently solve the explore-exploit dilemma (Gershman et al., 2018).

Under conditions of uncertainty, decision makers must decide not only *if* and *how* it is best to explore, but also *when* it is best to explore. In many instances, deciding when to explore is obvious when external changes in choice or reward contingencies introduce peaks in uncertainty about the current strategy. These peaks, in turn, signal that exploration would be beneficial. However, phasic exploration requires calculating the potential benefits of exploring each time uncertainty peaks, and thus can be a difficult or costly strategy to implement when there are additional sources of uncertainty (e.g. noisy stimulus-reward associations) that hinder our ability to correctly infer when exploration is warranted. In such instances it may be more advantageous for us to explore more regularly, consistently switching back and forth between decisions to explore or exploit. Tonic exploration, therefore, is more distributed throughout time but, like phasic exploration, is a form of directed exploration that remains focused on gathering information to reduce uncertainty and maximize reward.

Though it is largely assumed that directed exploration is phasic, driven by discrete periods of expected or unexpected uncertainty, there is evidence that tonic exploration is also a useful strategy when navigating changing environments. In a recent study by Ebitz et al. (2019), experimenters sought to identify a behavioral metric for tonic exploration. To do so, they analyzed the behavior of animals during an adapted version of the Conceptual Set-Shifting Task (CSST) to study prefrontal cognitive function in non-human primates (Moore et al., 2005). In the task, animals are presented with three visual stimuli, each with a unique shape and color. Of these six possible stimulus features (three shapes + three colors), one was randomly selected as the rewarding feature for each block. In other words, to earn reward, animals had to learn to respond only to stimuli that possessed a specific color or shape feature. For example, in a block where the rewarded feature was 'blue,' animals should choose the blue stimulus regardless of its shape to get a juice reward. Once animals made fifteen correct choices under a given feature rule, the rewarded feature was switched, and animals had to flexibly adapt their behavior to sample other options and discover the new rule.

Using this paradigm, Ebitz et al. (2019) identified two distinct types of errors that were relevant to exploratory behavior. The first was perseverative errors, which they define specifically as choices adhering to the *previously* correct rule that occurred within five trials of a rule change. Making fewer perseverative errors reflects increased behavioral flexibility in their task and suggests a more efficient use of phasic exploration when rules change. Secondly, they identified errors made in the ten trials *before* the most recent rule change as lapses. Typically, these kinds of errors are thought to reflect inattention or poor learning of some sort, but Ebitz et al. (2019) posit that they may instead be an expression of tonic exploratory noise in behavior. Consequently, instead of these two error types representing distinct and separate processes, the authors suggest that they may jointly reflect a broader, underlying exploratory drive that is upregulated during periods of high uncertainty (i.e. when the rules change) and downregulated during stable periods of the task.

When Ebitz et al. (2019) examined the relationship between lapses and perseverative errors, they found that the two were negatively correlated. Thus, in blocks where animals made more lapses during stable periods, they also tended to be more flexible and make fewer errors during periods of rapid change. This is the opposite of what we would expect if animals simply failed to learn, as that would lead to increased lapses as well as increased perseverative errors. Ebitz and colleagues (2019) also fit reinforcement learning models to quantify how much animals were using previous outcomes to update their current beliefs with a learning rate parameter. They found that, on average, blocks with higher lapse rates also had higher learning rates. In addition, they observed that perseverative errors in one block could not be explained by lapses in the preceding block, which would be expected if the perseverative errors were the result of poor learning in the previous block. Taken together, these results indicate that not all lapses during stable periods reflect a failure to learn, and instead some have an underlying exploratory cause that facilitates learning.

Interestingly, Ebitz et al. (2019) also found that chronic cocaine use altered the expression of tonic exploration. After assessing performance in the task at baseline, animals were trained to self-administer cocaine and subsequently tested during a period of chronic use.

Experimenters found that cocaine self-administration simultaneously increased perseverative errors and decreased the number of lapses made. Though both types of errors were affected, the negative correlation between them was not. In other words, the line of best fit for their relationship did not change in slope, it merely shifted along the axes. These results suggest that chronic cocaine use affects the common cause driving tonic and phasic exploration, implicating mechanisms of dopaminergic transmission in tonic exploration.

In addition to being the first to coin the use of ‘phasic’ and ‘tonic’ to describe exploratory states, Ebitz et al. (2019) establish a flexible framework for assessing tonic exploration. By demonstrating that seemingly random errors during stable periods of their task are meaningfully related to errors during critical periods of change and flexibility, their work implies that we may be prematurely dismissing lapses in choice behavior in other environments. They show evidence for tonic exploration in a very specific task environment where there is no randomness in choice feedback, only unexpected uncertainty in the form of rapid rule changes. As a result, the work of Ebitz et al. (2019) raises many questions about if and how tonic exploration exists in relation to other types of uncertainty. For example, in what other frameworks is tonic exploration useful? How might tonic exploration manifest as a response to different types of uncertainty?

To address these questions, we examine behavior during a reversal learning task where expected uncertainty is created by a single reversal in learned cue-reward relationships, as well as by variability in reward feedback. Importantly, we take a Bayesian approach to analyzing reversal learning and identifying exploration (Costa et al., 2015). Bayes’ theorem, at the core of all Bayesian analyses, is a model for learning from evidence. First described by Thomas Bayes in 1774, Bayes’ theorem has three central components (Finetti, 2017). The first, known as a prior, represents the prior knowledge we have about a given phenomenon before any observations are made. For example, most people would confidently assume that all of the coins in their wallet are 'fair' coins that have an equal chance of landing heads as tails when flipped. In that case, our prior belief for the probability a coin is fair would be 1. The second component of Bayes' theorem is the likelihood, or the probability that, given our prior beliefs, we observe the current evidence. So, for example, the likelihood of flipping heads if our belief about all coins being fair is true would be 0.5. The third component in Bayes' theorem is known as the posterior. The posterior integrates the prior and the likelihood, indicating the probability that our belief is true given the evidence we've observed. So, after flipping the coin ten times and having it come up heads every time, our posterior probability that coins are fair would decrease to be lower than 1 to reflect how this new evidence affected our beliefs about fair coins.

Recent work by Costa and colleagues (2015) exemplifies the advantage of a Bayesian approach to studying decision-making in reversal-learning tasks. In a reversal learning task where the relationships between two cues and their assigned reward values are programmed to switch halfway through each block, it follows that animals may develop a prior belief about the occurrence of a reversal as they become more experienced with the task. Further, if they have learned to expect an abrupt change in rules, they might use that knowledge to guide their behavior. Costa et al. (2015) test these intuitions by assigning a Bayesian prior to the period where animals may expect a reversal in cue-reward mappings. The likelihood, in this experiment, was the probability of observing the animals' choices (and the subsequent reward outcomes) given their prior belief about when a reversal would occur. With this information, Costa et al. (2015) calculate the animals' posterior belief that a reversal will occur on each trial, given the choice and reward data observed. Then, by assessing the distribution of this posterior belief across the span of each task block, Costa et al. (2015) are able to estimate the point at which animals make a switch in their choice behavior. Critically, this reversal in the animals' behavior does not necessarily correspond to the point in the task program where the cue-reward contingencies reverse. Instead, analyzing behavior with respect to this Bayesian estimate of the reversal point clarifies how the animal is accumulating evidence and using it to decide when to explore. In other words, the model put forward by Costa et al. (2015) allows us to flexibly analyze behavior with respect to the structure the animal infers from the task, rather than with respect to the rigid structure of the task design.

Lastly, we also incorporate analyses based on the matching law of behavior to clarify whether or not animals employ tonic exploration in our task. Matching law describes how, when selecting between multiple rewarding alternatives, animals tend to allocate their responses proportionally to the reinforcement associated with each alternative (Herrnstein, 1961). Across many primate and non-primate species, matching law has been shown to explain global choice behavior (de Villiers & Herrnstein, 1976; Lau & Glimcher, 2013; Pierce & Epling, 1983) However, actual choice often deviates from matching. For example, animals may choose high reward alternatives less than the matching law predicts. This deviation from matching, known as undermatching, has previously been interpreted as poor learning of the relative values of the two alternatives (Baum, 1974; Baum, 1979) or as randomness in the neural mechanisms that support decision-making (Soltani et al., 2006). In this way, choices that result from undermatching bear a resemblance to the lapse choices that Ebitz et al. (2019) observed during the stable periods of their task – both result in more frequent sampling of a less-rewarding alternative. Therefore, matching law provides an alternative framework for identifying and quantifying variation in decision-making that may complement existing analyses of tonic exploration.

Summary of Experiments

To determine whether animals strategically employ both tonic and phasic exploration during probabilistic learning, we reexamined monkeys' choice behavior during a two-arm bandit reversal learning task (Costa et al., 2015). Here, we leverage the same Bayesian analysis used by Costa et al. (2015) to quantify both the animals' belief that a reversal has occurred and their belief in how often choices will be rewarded – two types of information with the potential to generate uncertainty that facilitates exploration. Using these estimates, along with other task-relevant performance measures, we apply a similar framework to Ebitz et al. (2019) and examine lapses as a metric for tonic exploratory noise in behavior. Further, because the work of Ebitz et al. (2019) shows that the mechanism supporting tonic exploration may be regulated by dopaminergic systems, we examine the effects that two dopaminergic drugs – levodopa and haloperidol – have on exploration in the bandit task. Levodopa (L-dopa) is a precursor for dopamine that is commonly used to restore dopamine levels in patients with Parkinson's disease. When administered orally, it crosses the blood-brain barrier to be taken up by dopamine neurons, converted to dopamine, and released synaptically when those dopamine neurons are stimulated (Robinson et al, 2005). Haloperidol works by blocking D2 receptors, but it is not selective for the D2 receptor; it also has noradrenergic and cholinergic blocking action (Rahman & Marwaha, 2021). These drugs were chosen because they have been shown to exert opposing effects on reversal learning behavior (Cools et al., 2006; Cools et al., 2007; Frank & O'Reilly, 2006; Ridley et al., 1982; Shohamy et al., 2006), and may also have effects on directed exploration (Chakroun et al., 2020) that extend to the use of tonic exploration.

We show that while the use of phasic exploration and the number of lapses made in the reversal phase of our task depend heavily on the expected reward uncertainty, the use of tonic exploration does not. Additionally, neither phasic nor tonic exploration is affected by the administration of L-dopa or haloperidol. Moreover, the evidence we find for tonic exploration cannot be accounted for solely by the extent to which animals deviate from matching their responding to the reward rate of each option during the reversal phase of the task. Further, we find that lapses are predicted by trial-by-trial changes in the monkey's subjective beliefs about whether or not a reversal has occurred and whether or not a correct choice will be rewarded, indicating that tonic exploration is directed towards reducing task-related uncertainty.

Methods

All data used was collected and previously reported on by Costa et al., (2015).

Animals

Methods originally reported by Costa et al. (2015):

> "Three male rhesus monkeys (Macaca mulatta), aged 5– 6 years with weights ranging from 6.5 to 9.3 kg, were studied. All monkeys were placed on water control for the duration of the study and, on test days, earned all of their fluid through performance on the task. Experimental procedures for all monkeys were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the Animal Care and Use Committee of the National Institute of Mental Health."

Reversal learning task

Methods originally reported by Costa et al. (2015):

> "The monkeys completed 4–44 (20.93 +/- 0.93, mean +/- SE) blocks per session of a two-arm bandit problem. Each block consisted of 80 trials and involved a single reversal of the stimulus– reward contingencies (Fig. 1). On each trial, the monkeys had to first acquire and hold a central fixation point (250 –750 ms). After the monkey fixated for the required duration, two stimuli appeared to the left and right (6° visual angle) of the central fixation point. Stimuli varied in shape and color, and stimulus location (left vs right for each shape) was randomized within a block. Monkeys chose between stimuli by making a saccade to one of the two stimuli and fixating the cue for a minimum of 500 ms. One of the stimuli had a high reward probability, and one had a low reward probability. Juice rewards were probabilistically delivered at the end of each trial, followed by a fixed 1.5 s intertrial interval. A failure to acquire/hold central fixation or to make a choice within 750 ms resulted in a repeat of the previous trial. The three reward schedules used were 80/20%, 70/30%, and 60/40%. Use of these three reward schedules anticorrelates the mean reward probabilities of the bandit arms. The trial on which the cue–reward mapping reversed within each block was selected pseudorandomly from a uniform distribution across trials 30 –50. The reversal trial did not depend on the monkey reaching a performance criterion. Reward schedules were always constant within a block but could (and usually did) change across blocks."

Reward schedules that have two stimuli with a greater difference in reward probabilities (i.e. 80/20) are referred to here as 'easier' schedules for two reasons: First, because increasing the difference between the two reward probabilities assigned to the stimuli reduces the difficulty in determining which of the two stimuli is more frequently rewarded. Second, because increasing the discriminability between the assigned reward probabilities of the two stimuli also decreases the difficulty in detecting when the stimuli-reward mappings are reversed. Conversely, schedules that have two stimuli was a smaller difference in reward probabilities are referred to here as ‘harder’ or more difficult because of the decreased discriminability between the true reward probabilities of each stimuli.

> “Stimuli consisted of simple images of a circle and square in one of three colors (red, green, and blue). The two choice options always differed in color and shape. This resulted in six unique stimulus combinations. When these combinations were crossed with the three reward schedules and whether a particular shape was more or less initially rewarding (e.g., whether the blue square was the best choice before or after the reversal), this resulted in 36 block combinations. Block presentations were fully randomized without replacement. This ensured that a specific stimulus–reward combination was never repeated directly until all 36 block com- binations were experienced (< 4% of sessions). Although combinations were potentially repeated across sessions, during inspection, there was no evidence of improved performance across sessions.
>
> Each monkey received 10–14 d of initial training on the described reversal learning task until they were routinely completing 15–20 blocks per session. Animals first learned the structure of the task under a deter- ministic reward schedule. Probabilistic reward schedules were then in- troduced progressively until the animals exhibited stable performance on the tested reward schedules.
>
> Stimulus presentation and behavioral monitoring were controlled by a personal computer running the Monkeylogic (version 1.1) MATLAB toolbox (Asaad and Eskandar, 2008). Eye movements were monitored using an Arrington Viewpoint eye-tracking system (Arrington Research) and sampled at 1 kHz. Stimuli were displayed on an LCD monitor (1024 x 768 resolution) situated 40 cm from the monkey’s eyes. On rewarded trials, 0.085 ml of apple juice was delivered through a pressur- ized plastic tube gated by a computer-controlled solenoid valve (Mitz, 2005).”

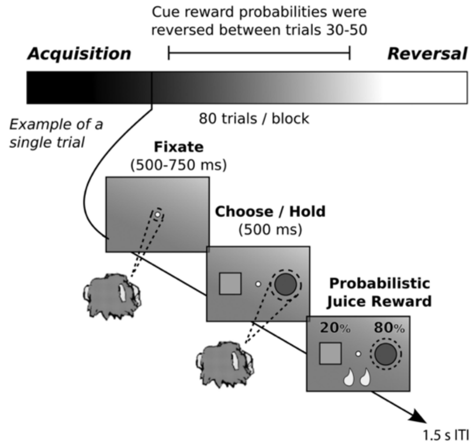
Cue reward probabilities were reversed between trials 30-50
Acquisition
Reversal
80 trials / block
Example of a single trial
Fixate (500-750 ms)
Choose / Hold (500 ms)
Probabilistic Juice Reward
20%
80%
1.5 s ITI

Figure 1: Block and trial structure in the two-arm bandit reversal learning task (taken from Costa et al., 2015). Each block contained 80 trials. The cue-reward mapping was reversed on a single trial randomly chosen between 30 and 50. Trials before the reversal are referred to as acquisition, and trials after the reversal are referred to as reversal. The reward schedule was always constant within a block (i.e., 80/20, 70/30, or 60/40%), but it usually changed across blocks. ITI = Intertrial interval.

Drug administration

Methods originally reported by Costa et al. (2015):

> "Before drug testing, monkeys were first habituated to intramuscular needle injections of saline given in conjunction with free juice (pH 7.4, 0.1 ml/kg). After this habituation period, monkeys readily presented their leg for injections. At the start of each placebo session—while chaired and outside of the test box—the monkeys received an intramuscular injection of saline (1 ml) while they drank 6 ml of apple juice from a plastic syringe. They were then head posted and placed inside the test box. The eye-tracking system was then calibrated to avoid drug-related effects on eye-tracking sensitivity. During the remainder of the wait period, the animals viewed a nature movie. This placebo procedure was consistent with the two methods of drug administration. Free juice was similarly delivered at the start of each drug session before waiting 30 min to start the task.
>
> On days the monkeys received L-DOPA, we dissolved, under sonication, a pulverized fixed dose tablet of L-DOPA (100 mg/25 mg carbidopa; Actavis) into the delivered free juice and paired it with an intramuscular injection of saline. On days the monkeys received haloperidol, free juice was delivered in conjunction with an intramuscular injection of haloperidol (6.5 μg/kg; Bedford Laboratoies). This dose was consistent with doses shown previously to have behavioral effects (Turchi et al., 2010). Injections were prepared by first dissolving a fixed dose of haloperidol (100 μg) under sonication into PBS under strict sterile conditions and stored at 4°C for use within the week. On the day of the drug injections, aliquots were resonicated and allowed to reach room temperature before injection. Injections were given intramuscularly into the lateral hindlimb.
>
> The monkeys completed multiple sessions under each drug condition. On L-DOPA, monkey E completed seven sessions comprising 138 total blocks, monkey G completed six sessions comprising 159 total blocks, and monkey M completed seven sessions comprising 193 total blocks. On haloperidol, monkey E completed seven sessions comprising 100 total blocks, monkey G completed eight sessions comprising 142 total blocks, and monkey M completed seven sessions comprising 143 total blocks. The total number of placebo sessions ranged from 15 to 24 sessions per animal (22 for E, 24 for G, and 16 for M), comprising 370 – 479 blocks. Haloperidol sessions were spaced a minimum of 7d apart to facilitate washout, whereas the faster clearance of L-DOPA permitted a minimum spacing of 3d between sessions. L-DOPA and haloperidol sessions were interleaved and counterbalanced for the day of the week to minimize routine caretaking effects on behavior. Each drug session was preceded by at least one placebo session, and all placebo sessions lagged the most recent drug session by a minimum of 2d to minimize carryover effects."

Bayesian model

The Bayesian model was developed and originally implemented in MATLAB by Dr. Vincent Costa and Dr. Bruno Averbeck (Averbeck & Costa, 2017; Costa et al., 2015). Here, the generation of Bayesian posterior distributions and estimation of reversal points represents work done to confirm the prior model findings by adapting the code for Python.

Briefly, the goal of the Bayesian modeling approach here was to quantify the amount of posterior evidence animals had that 1) a reversal occurred or 2) that their choices would be rewarded, given their choice and outcome histories. The model by Costa et al. (2015) can do this from the perspective of an ideal observer who has information about both choice and reward histories, and therefore whose posterior estimate of a reversal or reward reflects all of the evidence that was available to the animal. The model can also do this from the perspective of the animal, using only information about their history of choices, to get a posterior estimate that is more reflective of their subjective beliefs. Posterior estimates are calculated for each trial from 0 to 80, resulting in a posterior distribution that reflects the changes in evidence (or belief) as they evolve throughout each block of the task. Using these posterior distributions, we can also use the model to estimate the specific point at which the ideal observer or animal switched their choice preference. For clarity on the structure of the model, the original methods from the Costa et al. (2015) manuscript are copied below:

> "We fit three Bayesian models that estimated the posterior probability that reversals occurred on each trial, under various assumptions. To estimate the models, we fit a likelihood function given by the following:
>
> $$f(x, y|r, p, h, M) = \prod_{k-1}^{T} \mathrm{q(k)} \tag{1}$$
>
> where r is the trial on which the reward mapping reversed ($r \in 0 - 81$), and p is the probability of reward for the high reward option (models 1 and 3) or the consistency with which the animals chose their preferred option (model 2). The variable h encodes

whether option 1 or option 2 begins the block as the high reward option ($h \in 1, 2$), k indexes trial number in the block, and T is the current trial. The variable r ranges from 0 to 81 because we allowed the model to assume that reversals occurred before the block started or after the block ended. In either of these cases, there would be no switch within the block (the model estimated no reversal in <1% of the total blocks analyzed), and the posterior probability of a switch would be equally weighted for r equal to 0 or 81. The data are given by the vectors x and y, where the elements of x are the rewards ($x_i \in 0, 1$), and the elements of y are the choices ($y_i \in 1, 2$) in trial i. We fit three variants of this model indicated by M ($M \in 1, 2, 3$). $M = 1$ is the ideal observer. This model was used to estimate the evidence the animal had available to it when it made its decisions, as well as the ideal reversal point. $M = 2$ is the behavioral choice model. This model was used to estimate where the animal reversed its choice behavior." The third model, $M = 3$, is a causal version of the ideal observer model ($M = 1$).

"The behavioral choice model ($M = 2$) estimates the trial on which the animals switched their choice behavior. This only depends on the pattern of choices, not on whether they were rewarded. We assumed that the animal had a stable choice preference that switched at some point in the block from one stimulus to the other. Given the choice preference, the animals occasionally chose the wrong stimulus (i.e., the stimulus inconsistent with their choice preference) at some lapse rate 1 - p. Thus, for $k < r$ and $h = 1$, choosing option 1, $q(k) = p$; and choosing option 2, $q(k) = p$. For $k = r$ and $h = 1$, choosing option 1, $q(k) = 1 = p$; and choosing option 2, $q(k) = p$. Correspondingly, for $k = r$ and $h = 2$, choosing option 2, $q(k) = p$, etc. Thus, this model assumed that the monkey preferred one option before switching and preferred the other option after switching. It most often chose its preferred option ($p < 0.5$), but it occasionally chose the wrong target perhaps as a result of lapses in attention. For all reported analyses, we marginalized over the correct choice rate p. Therefore, we assumed that the animals were maximizing and not doing probability matching. These values for $q(k)$ were filled in for the entire block, because we were performing this analysis post hoc to estimate where the animal reversed.

For models 1 and 3, we estimated whether a reversal had occurred conditioning only on outcomes before the current trial, T. This provided an estimate of the information on which the animal was making its choice. For these models, values of $q(k)$ for each schedule were given by the following mappings from choices to outcomes. For $k < r$ and $h = 1$ (before reversal and target 1 is the high probability target), choose 1 and get rewarded $q(k) = p$; choose 1 and not get rewarded, $q(k) = 1 - p$; choose 2 and get rewarded, $q(k) = 1 - p$; and choose 2 and not get rewarded, $q(k) = p$. For $k \geq r$, these probabilities flip. Correspondingly, for $k < r$ and $h = 2$, the probabilities are also complimented. These values were filled in [for the entire block for model 1, and] up to the current trial, T [for model 3].

Given these mappings for q(k), we could then calculate the likelihood as a function of r, p, and h for each block of trials. The posterior is given by the following:

$$p(r,p,h|x,y,M) = \frac{f(x,y|r,p,h,M)p(r|M)p(p,h|M)}{p(x,y|M)} \tag{2}$$

The priors on p, h, and r were flat for all models. Given the priors, the posterior over switch trial could be calculated by marginalizing over p and h. Specifically,

$$p(r|x, y, M) = \sum\nolimits_{p,h} p(r, p, h|x, y, M) \tag{3}$$

Similarly, the posterior over the probability of reward for the high probability option could be calculated by marginalizing over r and h:

$$p(p|x, y, M) = \sum\nolimits_{r,h} p(r, p, h|x, y, M) \tag{4}$$

After the posterior over r for both models was calculated, the expected reversal point was calculated as:

$$< r|M >= \sum_{r=0}^{81} r * p(r, p, h|x, y, M) \tag{5}$$

Because the estimated reversal point was not guaranteed to be an integer, it was rounded to the nearest integer when it served as an index of summation. Trials less than $\{r|M\}$ were assigned to the acquisition phase, whereas trials greater than or equal to $\{r|M\}$ were assigned to the reversal phase."

Analysis of phasic and tonic exploration

We quantified phasic exploration as the signed deviation of the monkey's reversal point (M=2) from that of the ideal observer (M=1). For each model, the reversal point in a given block was calculated as the weighted mean of the posterior probability of reversal across all 80 trials, as expressed in equation (5) above. Because the estimated reversal point was not guaranteed to be a whole number, it was rounded to the nearest whole number after summation. We then compared the estimated reversal points of the ideal observer (M=1) and behavioral choice (M=2) models to capture how monkeys performed relative to an observer with full knowledge of the environment. Because the ideal observer model captures the full scope of evidence that was available to the animal (Costa et al., 2015), comparing the animal's reversal point to that of the ideal observer gives us a proxy for how much or how little evidence the animal accumulated before reversing its choice behavior, and thus is indicative of how the animal was using phasic exploration.

We quantified tonic exploration as the rate of lapses (choices of the low probability option) in the last twenty trials of each block. We chose this period, specifically, because at trial 50 in the task program a reversal *must* have occurred. After this point, cue-reward mappings are stable and monkeys (or the ideal observer) have sufficient information to infer a reversal. For this reason, blocks in which the estimated reversal point in the monkeys' behavior was past trial 60 were excluded from analyses (<1% of all blocks). Values were calculated for each block and then averaged in each session separately for each reward schedule.

Deviation from matching behavior

For choices in the reversal phase, we calculated the rate at which monkeys deviated from matching to the high value option as (Reed & Kaplan, 2011):

$$DM_{HV} = P(chosen)_{HV} - \frac{RewardRate_{HV}}{RewardRate_{HV} + \ RewardRate_{LV}} \tag{6}$$

where fractions ($RewardRate_{HV}$, $RewardRate_{LV}$), the rewards rates for the high and low value options, are defined as:

$$RewardRate = \ AssignedRewardRate * p(chosen) \tag{7}$$

Statistical analyses

Statistical analyses employed mixed-model ANOVAs, carried out in JMP 14 (SAS). Drug, schedule, and monkey were all specified as fixed effects with interactions, while session (nested under monkey) was specified as a random effect. Dependent variables (in separate models) included the signed difference in reversal points, the absolute difference in reversal points, and lapse rates. Post hoc analyses of significant main effects used Tukey's HSD test to simultaneously test all pairwise comparisons while controlling for the family-wise error rate of multiple comparisons.

A hierarchical linear model was used to estimate the effects of posterior beliefs on the animals' choice behavior, and was also carried out in JMP 14. We first ran fit logistic regression with choice behavior on a given trial as the dependent variable (lapse/no lapse) reward on previous trial, posterior on schedule, and posterior on reversal as predictors by each schedule*drug*monkey combination. Then, we extracted the beta weights for each of these predictors in the models and used them as observations for a mixed-effects ANOVA model using schedule, and drug as fixed effects and monkey as a random effect.

Pearson's correlations were carried out in Python 3.7.10 using the Scipy library. Correlations were used to examine the relationship between the signed difference in reversal points and lapse rates. Partial correlations were used to further examine the relationship between the signed difference in reversal points with either lapse rate or deviation from matching independently while accounting for the variance explained by the other. To do this, we fit an ANOVA model in JMP 14 (SAS) using monkey drug, lapse rates and deviation from matching as fixed effects and session (nested under Monkey) as a random effect to predict the signed difference in reversal points. Then we reconstructed the expected values for the signed difference in reversal points using the observed values, model intercept and model beta values for either lapse rate or deviation from matching.

Results

Use of phasic exploration in a two-armed bandit task reversal learning task

We replicated the Bayesian modeling approach outlined by Costa et al. (2015) to calculate the probability of detection of a reversal based on both an ideal observer (M=1) and on monkeys' choice behavior (M=2) in every block. From these distributions, the model allowed us to estimate the trial on which the ideal observer (M=1) and the monkey (M=2) reversed their behavior (see Methods). First, we replicated analyses of schedule-related effects on the average posterior probability distribution for reversals in the monkey's (M=2, Fig. 2a) and the ideal observer's (M=1, Fig. 2b) modeled behavior. This replication confirmed the finding of Costa et al. (2015), using new code adapted for Python 3.7.10, that posterior probability distributions were broader and had a larger left tail in more difficult schedules, reflecting the increased difficulty of detecting reversals when reward uncertainty was higher.

We also replicated analyses done by Costa et al. (2015) to examine whether reward schedule affected the amount of evidence necessary to trigger a reversal. To do this, Costa et al. (2015) re-aligned the posterior evidence of the ideal observer (M=1) that a reversal occurred to the trials surrounding the monkey's reversal (Fig. 2c). As was found by Costa et al. (2015), we saw in our replication that reversal points in the monkey's choices clearly followed peaks in the posterior distribution (Fig. 2c). As the reward schedule became more uncertain, monkeys' reversals followed smaller peaks in the posterior distribution (schedule, $F_{(2,138)} = 931.70$, $p < 0.001$), confirming that evidence for reversal scales negatively with the uncertainty in reward feedback.

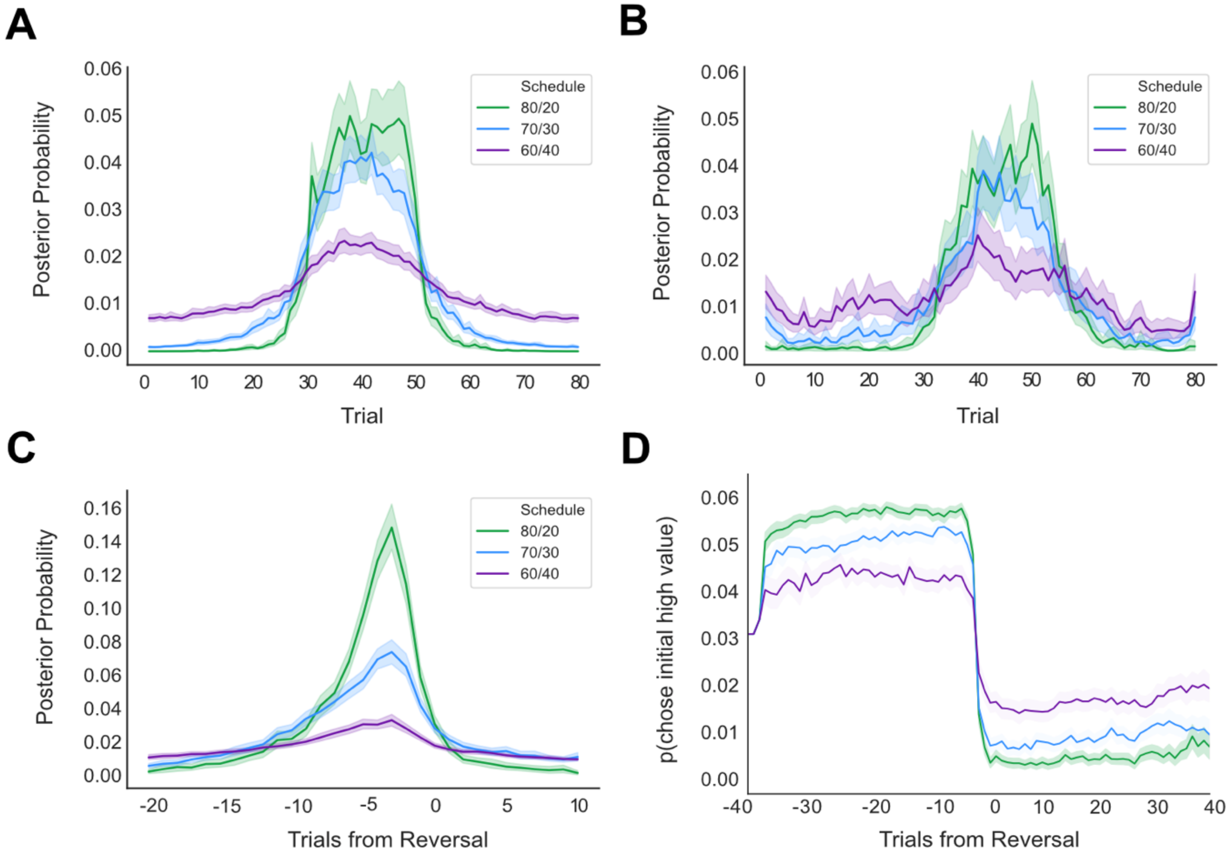

Figure 2: Bayesian posteriors on reversal by reward schedule. Shading indicates one standard error of the mean (SEM), and grey bars indicate the trial range in which a reversal could have occurred. (A) Mean posterior probability on reversal for the ideal observer model (M = 1) for each reward schedule. (B) Mean posterior probability on reversal for the behavioral choice model (M = 2) for each reward schedule. (C) Mean posterior probability on reversal for the ideal observer model (M = 1) by schedule, aligned to the estimated trial on which the animals reversed their choice behavior (always trial = 0 here). (D) Summary of choice behavior in the task, aligned to the estimated trial on which the animals reversed their choice behavior (always trial = 0 here).

We then sought to quantify the monkeys' use of phasic exploration (i.e. reversing their choice behavior during the window where the benefit of exploring was greatest). We assumed the monkeys had learned that there would be a single reversal of the reward contingencies, and Costa et al. (2015) previously reported that they try to anticipate the contingency reversals. With full knowledge of the environment (M=1), an ideal observer can calculate the benefit of switching choice strategy and do so when uncertainty peaks. By looking at how closely the monkey's reversal in choice behavior (M=2) is aligned to that of the ideal observer (M=1), then, we can capture the extent to which the monkey is successfully using phasic exploration. Analysis of the signed reversal points revealed that monkeys reversed their behavior earlier, with respect to the ideal observer, on average ($t_{(80)} = -4.79$, $p < 0.001$), and that this effect varied across the three reward schedules (Schedule, $F_{(2,169)} = 53.35$, $p < 0.001$). On average and in reference to the ideal observer (Fig. 3a), the monkeys reversed their behavior earlier during 60/40% blocks compared with 70/30% ($t_{(89)} = 4.77$, $p < 0.001$) or 80/20% blocks ($t_{(89)} = 8.77$, $p < 0.001$). In blocks where animals received L-dopa or haloperidol, there were no effects of drug on the signed difference in reversal points between the animal and the ideal observer (Fig. 3b).

To confirm that phasic exploration involves monkeys triggering off of accumulating evidence, we correlated the signed difference in reversal points with the cumulative sum of the ideal observer's posterior (Fig. 3c). Intuitively, we found a positive correlation between the signed difference in reversal points and the accumulated evidence at the monkey's reversal point at the session level ($r_{(74)} = 0.75$, se = 0.018) illustrating that an increased readiness to employ phasic exploration corresponds with reversing behavior on the basis of smaller evidence peaks.

We also correlated the cumulative sum of the ideal observer's posterior with the absolute difference in reversal points (Fig. 3d) and observed a negative correlation across schedules ($t_{(273)} = -2.174$, $p = 0.031$), reinforcing the idea that animals were able to reverse more accurately to ideal when that reversal was based off of more evidence. With respect to drug, animals given haloperidol showed a significantly less negative correlation between evidence accumulated and absolute difference in reversal point than animals given L-dopa ($t_{(85)} = 2.56$, $p = 0.006$) or saline ($t_{(87)} = 2.74$, $p = 0.004$). This effect replicated the finding from Costa et al (2015) that haloperidol strengthens the monkeys' prior on when a reversal will occur. Here, however, we are using evidence accumulation (rather than a set assumption on the prior) to assay the model's belief a reversal has occurred (see Methods).

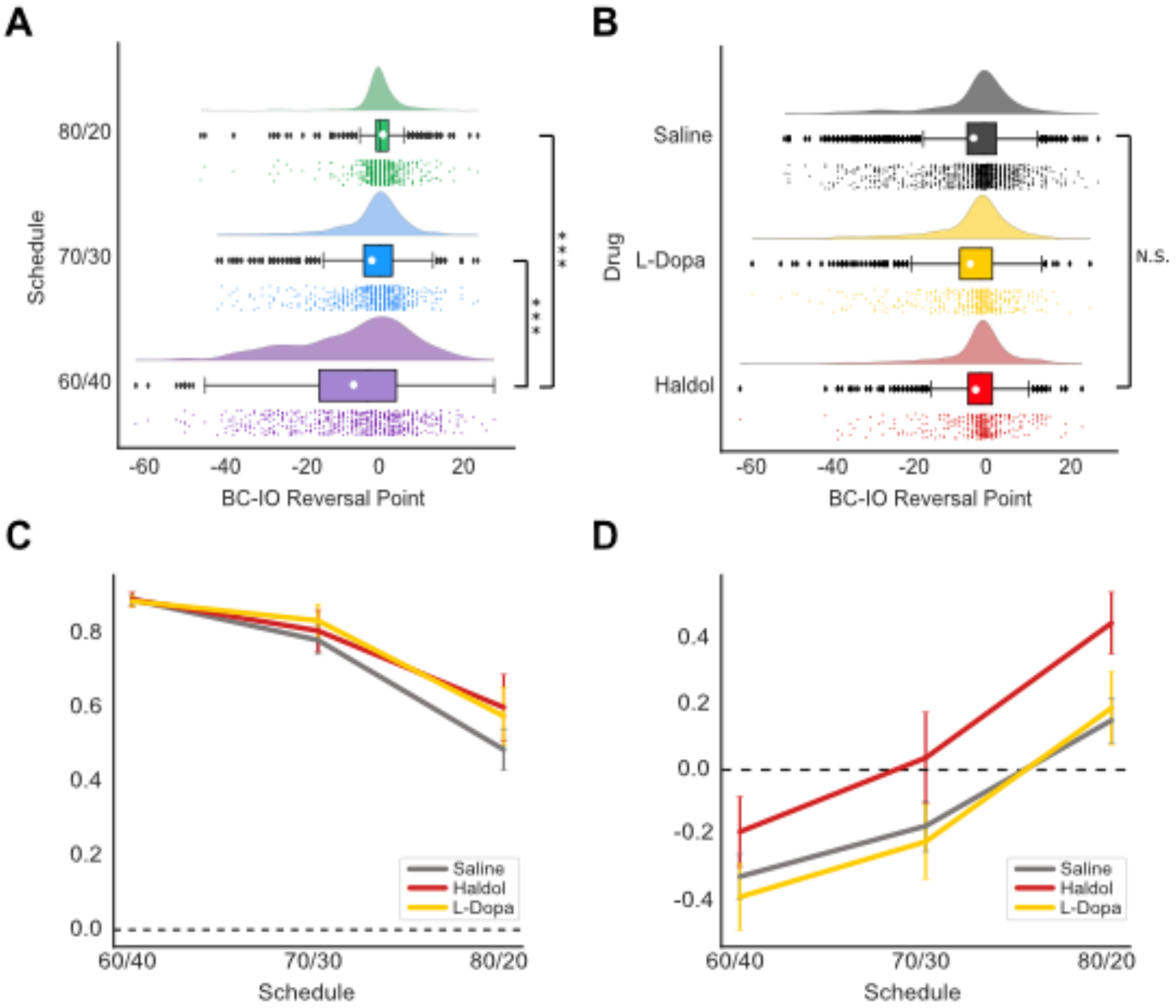

Figure 3: Reward uncertainty, not drug, affects the use of phasic exploration. (A) Dots represent individual data points, white circle on each boxplot indicates the mean value. The signed difference in the model-estimated reversal points for the behavioral choice (M=2) and the ideal observer (M=1) models, averaged by (A) schedule or (B) drug. (C) Error bars represent one SEM. Correlation (r-value) between the amount of evidence accumulated by the ideal observer at the animal's reversal point and (C) the signed difference in reversal points or (D) the absolute difference in reversal points.

Lapses, as a behavioral metric, quantify the use of a tonic exploration strategy

First, we examined the monkeys' choice behavior after trial 60 to determine how often they were lapsing during the stable period of the reversal phase (see Methods). There was a significant effect of reward schedule ($F_{(2,135)} = 150.18$, $p < 0.001$) on the amount of lapses made (Fig. 4a). Post-hoc comparisons showed that monkeys generally had lower lapse rates in the 80/20 schedule than in the 70/30 schedule ($t_{(129)} = -7.17$, $p < 0.0001$) and in the 60/40 schedule ($t_{(129)} = -17.29$, $p < 0.0001$). There was no effect of drug on the number or rate of lapse choices within sessions (Fig. 4b). Broadly, these comparisons indicate schedule-dependent lapses in task performance that scale positively with reward uncertainty.

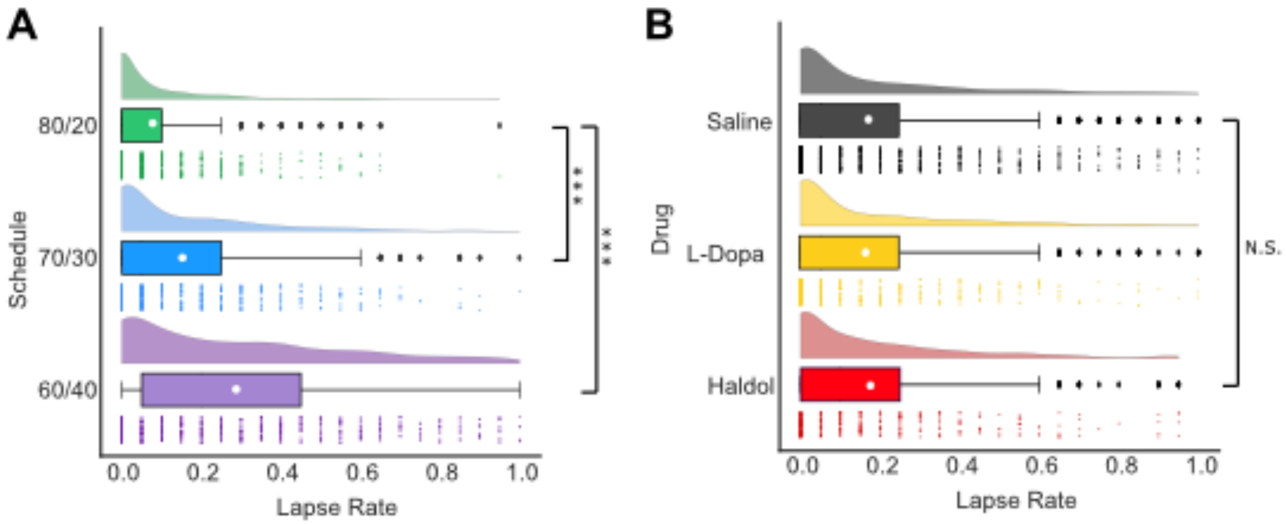

Figure 4: Reward uncertainty, not drug, affects lapse rates. (A) Dots represent individual data points, white circle on each boxplot indicates the mean value. The number of lapses made in the last twenty trials of each block, averaged by (A) schedule or (B) drug.

In order to establish that lapses, in our task, are an appropriate measure of tonic exploration, we sought to determine whether or not they were related to the signed difference in reversal points - a measure of phasic exploration. If lapses do reflect a form of tonic exploration driven by the same process as phasic exploration, we would expect a negative correlation between the two measures (i.e. more lapses as the animals reverse earlier with respect to the ideal and vice versa). If lapses are, on the other hand, some nuisance process, we would not expect them to be related (or positively correlated with) the difference in reversal points (Fig. 5a).

Because lapses could also result from undermatching, the tendency to randomly allocate responses to a choice alternative without regard for the rate of reinforcement (Wearden, 1983) we examined the correlation between lapses and phasic exploration while accounting for choice variability due to undermatching during the entire reversal phase. Undermatching was inversely correlated with lapse rates in a schedule dependent manner (Fig 5b; $r_{(60/40)} = -0.48$, $r_{(70/30)} = -0.85$, $r_{(80/20)} = -0.89$), all $p < 0.001$) and did predict reversal learning performance (Fig. 5c). Even so, lapse rates over the last 20 trials were still predictive of the relative amount of evidence acquired at the time the monkeys' decided to switch their choice preference (Fig 5d; ($F_{(1, 446)} = 35.53$, $p < 0.001$) and this effect did not vary by reward schedule. Not only does this suggest that lapses reflect tonic exploratory noise in behavior, it indicates that the use of tonic exploration is not dependent on reward uncertainty.

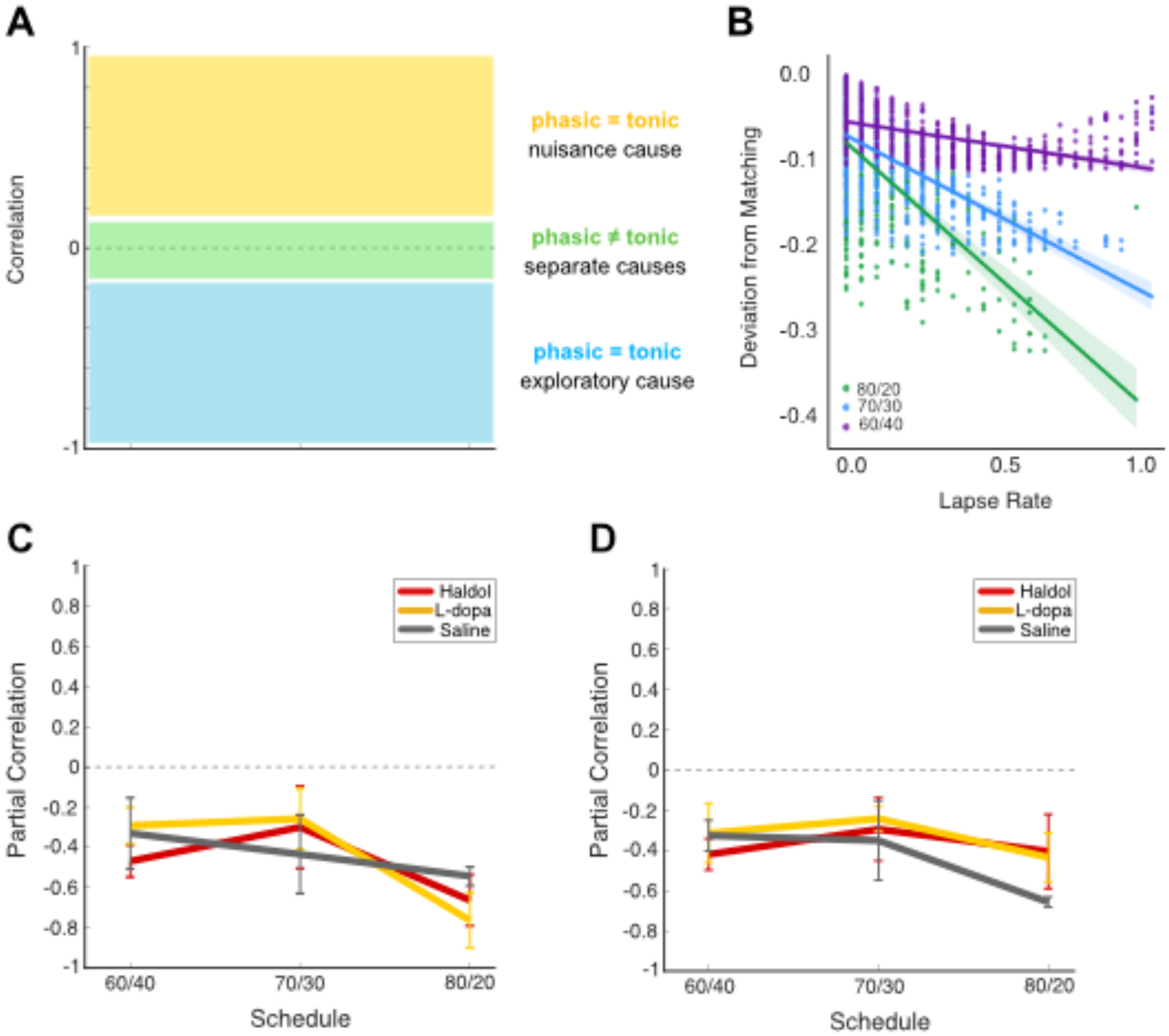

Figure 5: Evidence that lapses have an underlying exploratory cause. (A) Schematic adopted from Ebitz et al. (2019) depicting the possible relationships between lapses and the signed difference in reversal points. (B) Relationship between calculated deviation from matching and lapse rate in the last twenty trials of each block, shown by reward schedule. Dots represent individual data points, with regression lines plotted. Shaded area represents 95% confidence interval for the regression line. (C) Partial correlation, shown by schedule and drug, the signed difference in reversal points and (C) lapse rates or (D) deviation from matching, each after accounting for the variance explained by the other.

Lapses serve as a directed, not random, form of exploration

To test whether tonic exploration was driven by changes in the monkey's posterior beliefs, we implemented a logistic regression analysis to predict when the monkeys would make lapses over the last 20 trials of each block. We determined if we could predict lapses using the posterior probability that a reversal had occurred and the subjective reward rate (i.e. how consistently the monkeys' choice of the correct option was rewarded) calculated up until the immediately preceding trial. The posterior probability on reversal for the current trial ($\beta = 1.04$, $t_{(8)} = 7.25$, $p < 0.001$) and the subjective reward rate for the current trial ($\beta = -1.44$, $t_{(8)} = -12.64$, $p < 0.001$) both predicted lapses on the subsequent trial (Fig. 6). In addition, both of these estimates predicted lapses such that they were inversely related ($F_{(1,2)} = 71.00$, $p = 0.014$). The consistent relationship between these different posterior beliefs and lapses indicate that lapses do not occur randomly in time. Instead, our findings suggest that lapses track transient peaks in uncertainty during the stable period of the task, acting as a directed form of exploration.

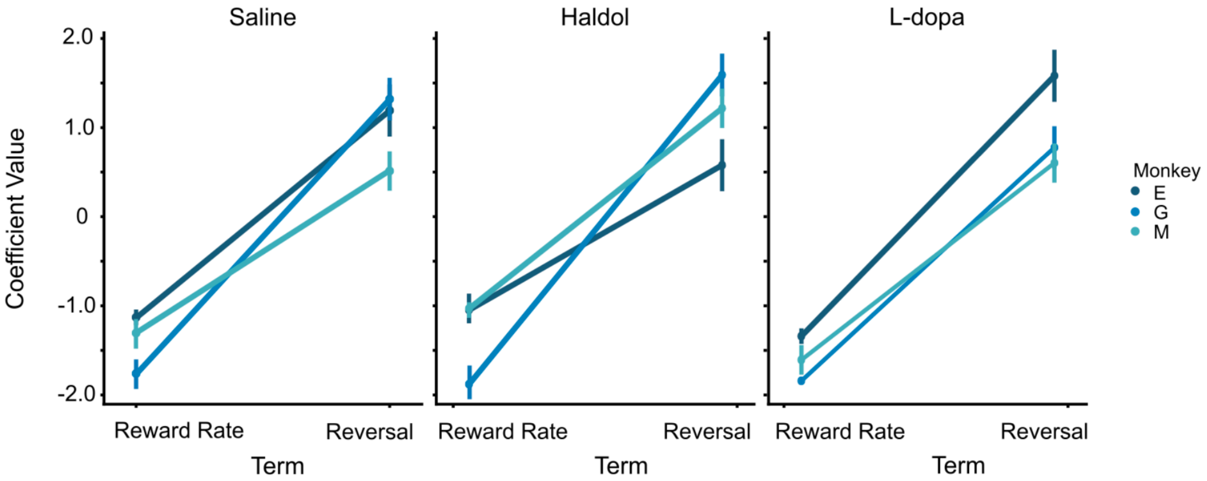

Figure 6: Lapses are predicted by beliefs about expected uncertainty. Coefficient values from a logistic regression using, on a given trial, the posterior probabilities (M=2) that a reversal occurred and that a choice would be rewarded to predict lapse choices on the subsequent trial. Shown by drug and animal. Error bars represent one SEM.

Discussion

Here, we examined the effects of expected uncertainty on different types of exploratory behavior in a probabilistic reward environment. We established that the use of phasic exploration, operationalized as the difference between the animal's reversal point and that of an ideal observer, varied by reward schedule. In easier schedules, with less reward uncertainty, animals reversed more closely to the ideal reversal point. In harder schedules, with more reward uncertainty, animals tended to reverse significantly earlier than ideal. Lapses in choice performance were also sensitive to reward uncertainty, with animals lapsing less in easier schedules and more in harder schedules. The use of phasic exploration and the lapse rate were both unaffected by the administration of L-dopa or haloperidol. We showed that lapses were negatively correlated with reversal accuracy, suggesting that they share some common exploratory cause (Ebitz et al., 2019). Further, this correlation was consistent across all three probabilistic reward schedules, indicating that tonic exploration occurs independently of the level of expected uncertainty animals have with respect to the value of their choices. Lastly, we found that lapses were predicted by the animals' subjective beliefs about expected uncertainty in their environment. Specifically, animals were more likely to lapse when their belief that a reversal had occurred was high and their estimation of the reward rate was low, suggesting that lapses are directed towards reducing moment-to-moment uncertainty. Taken together, these results show that lapses reflect a form of directed, tonic exploratory noise in response to expected uncertainty.

Our results build off of work by Ebitz et al. (2019), who first demonstrated that lapses can represent a form of tonic exploratory noise. The reward feedback in the task used by Ebitz et al. (2019) was deterministic, meaning that it had no variation and therefore was fully informative about whether or not each choice was correct. As a result, the only uncertainty in the task stemmed from the unsignaled rule changes. While this was likely an unexpected source of uncertainty at first, animals were trained to criterion on the task and continued to perform it throughout a two-month period of chronic cocaine self-administration. Therefore, the animals likely had a model of the task structure that includes rule changes as a form of expected uncertainty. In this kind of environment, where rule changes are expected and occur fairly often, tonic exploration is a useful strategy because monkeys can regularly make information-gathering choices instead of calculating the potential benefit of phasic exploration on every trial.

In our task, expected uncertainty is also created by a rule change, however there is only a single rule change that animals learn to anticipate in the middle of the block (Costa et al., 2015). As a result, animals do not necessarily need tonic exploration to succeed. Rather, they could simply accumulate evidence from the outcomes of their choices and employ phasic exploration when enough evidence accumulates that their current preferred choice is no longer the best one. If that was the case, and tonic exploration was not present, we would not expect to see any lapses in the reversal period of the task. Instead, what we see is that animals continue to make exploratory lapses well after they have employed phasic exploration. These data support the notion that tonic exploration is inherently coupled to the same process drive that supports phasic exploration and is a more general phenomenon that extends across paradigms. Surprisingly, given that lapse rates were highly affected by reward schedule in our task, the negative correlation between lapses and reversal accuracy was not schedule-dependent. Thus, tonic exploration seems to be insensitive to changing levels of expected uncertainty. Distinctions between tonic and phasic exploration and how they respond differently to unexpected versus expected uncertainty should be further examined to better understand how animals are able to implement these complex exploratory strategies.

Our results are consistent with those of Ebitz et al. (2019) in that we find tonic exploration to be directed towards the reduction of uncertainty. However, where Ebitz et al. (2019) find the timing of lapses to be random with respect to the task, we find that they are predicted moment-to-moment by animals' subjective beliefs about environmental uncertainty. One key advantage of our Bayesian modeling approach here is that it enables us to estimate, on a trial-by-trial basis, the monkeys' subjective beliefs with respect to two different aspects of expected uncertainty: whether or not a reversal in reward contingencies has occurred, and the reward rate for a chosen option. We find that both of these estimates on a given trial jointly predict whether or not animals will lapse on the subsequent trial. Thus, our results suggest that tonic exploration is directed towards gathering information when estimates of expected uncertainty fluctuate. Because we find the timing of lapses to be less random than Ebitz et al. (2019), it is possible that phasic and tonic exploration are more akin to the ends of a continuous spectrum rather than distinct processes. In this way, differing demands of expected or unexpected uncertainty may regulate the extent to which tonic exploration is spontaneous, compared to phasic exploration which is highly dependent on external constraints for its timing. Future work investigating the question of *when* to explore should continue to probe the continuity between phasic and tonic exploration, which may provide insight into the common mechanism(s) supporting them.

Expected uncertainty, volatility, and unpredictability

While we were framing phasic and tonic exploration in terms of expected uncertainty in our analysis, other models of decision-making in similar paradigms have proposed a hierarchical model of uncertainty that shapes dynamic learning, and that may be relevant here (Behrens et al., 2007; Piray & Daw, 2020). In these models, volatility reflects the rate of change (or noise) in the true cue values while unpredictability reflects how noisy the outcomes of choosing those cues are (Piray & Daw, 2020). The expected uncertainty in our task is due to both volatility and unpredictability; we can think of the stochasticity in reward outcomes as generating unpredictability, while the reversal in cue-outcome contingencies generates volatility. Our Bayesian model's estimate of the subjective reward rate (the probability a given choice will be rewarded) provides a rough approximation of the perceived unpredictability. Our model's estimate of whether or not a reversal occurred, however, does not directly capture volatility as it is defined in these models: it is an inference about whether the state of the environment has changed, not the rate at which the environment is changing. A clear next step is to manipulate unpredictability and volatility independently with respect to tonic exploration and observe how these types of uncertainty affect the computations we believe to underlie phasic and tonic exploration. For example, it is possible that tonic exploration is more useful in volatile environments as it allows decision-makers to maintain the flexibility needed to locate a constantly changing best option. Phasic exploration, on the other hand, might be better used in unpredictable environments to gate exploration by a higher threshold of uncertainty.

Application of the matching law

It is important to note the differences between our task and those traditionally used to study matching behavior when interpreting the deviation from matching scores in our results. The matching law was originally developed and studied using concurrent schedules of reinforcement for two alternatives, each delivering reinforcement on a variable interval schedule (Baum, 1974; Herrnstein, 1961). The use of 'concurrent' schedules simply indicates that the schedule for one alternative is independent of the schedule for the other. In our task, even though the probabilities of reward are related in that the two will always sum to one-hundred, the actual delivery of reward on each alternative is independent of the other. One could argue that the probabilistic delivery of reward in our task resembles either a variable interval (VI) schedule – where reinforcement is given after a specific, but varying amount of time – or a variable ratio (VR) schedule – where reinforcement is given after a specific, but varying number of responses (Zuriff, 1970). Because our task program uses varying probabilities to determine when reward is given, it closely resembles the varying ratios in a VR schedule. However, previous work has shown that in concurrent VR schedules, behavior closely adheres to what matching law predicts (Herrnstein & Loveland, 1975), which is not what we observe here. This may be because the rigid trial structure in our task (i.e. fixed intervals set for how long the animal has to respond, how long they need to fixate on a cue to 'select' it, etc.) imposes temporal structure that is not present in traditional operant paradigms used to study matching behavior. Typically, these paradigms use a 'free-operant' setup where one stimulus is continuously available for operant responses (Morris, 1987), unlike the trials in our task where stimuli are available for responding during discrete periods of time. As a result, our task may occupy a space somewhere in between VR and VI reinforcement schedules that makes application of the matching law less straightforward.

Another departure from traditional matching behavior analyses in the work here is our calculation of the experienced reward rates. Most commonly, the Generalized Matching Equation (GME) is used to calculate how much animals are deviating from matching behavior (Baum, 1974). In this equation, which compares the relative rates of response to the relative rates of reinforcement for two alternatives, it is always the *obtained* rates of each that are used. In other words, the GME uses the value of how much reinforcement was actually obtained by an animal when they chose one alternative instead of the assigned probabilistic rate of reinforcement for that alternative (Baum, 1974). In our calculation of scores for the deviation from matching, we calculate reinforcement for a given alternative as the product of the assigned reinforcement rate of that alternative and the probability that the animal chose it. While this value will eventually come to approximate the experienced reward rate in large datasets with many trials, like ours, it is still not an exact calculation of the experienced reward and therefore may deviate from other calculations of undermatching.

More recent work has also emphasized the ways in which matching behavior evolves on smaller, compared to larger or more global, timescales (Iigaya et al., 2019; Trepka et al., 2021). Based on the overarching hypothesis that matching behavior on a larger timescale develops from smaller adjustments of behavior based on trial-by-trial choices and outcomes (Trepka et al., 2021), we may also want to evaluate matching alongside lapse rates across different timescales in our task. For example, future work could expand or shrink the window in which we examine lapses during the task to see if lapses correspond with more local or more global undermatching behavior.

Neuromodulators and tonic exploration

Broad evidence suggests that dopamine helps regulate the explore-exploit tradeoff. Midbrain dopamine encodes reward prediction errors (Schultz & Dicksinson, 2000), which in turn help to encode the representations of cue value or action value (Collins & Frank, 2016; Flagel et al., 2011; Hamid et al., 2016) that guide explore-exploit decisions. However, because dopamine is also involved in processes besides learning such as energy expenditure (Beeler et al., 2012; Salamone et al., 2005), risk-taking behavior (Stopper et al., 2014), and motivation (Niv et al., 2007), parsing its precise role in exploration proves difficult.

Increasingly, work in the field of decision-making points to a role for dopamine signaling in directed exploration. For example, the systemic administration of L-dopa, putatively increasing tonic dopamine levels, has been shown to reduce directed but not random exploration by weakening neural representations of uncertainty (Chakroun et al., 2020). Interestingly, in the same study, the systemic administration of haloperidol had no effects on directed or random exploration. There is also evidence that dopamine mediates the signaling of an 'uncertainty bonus' that biases humans or animals to choose novel, less certain options (Costa et al.. 2014; Costa et al., 2019; Frank et al., 2009; Gershman & Tzovaras, 2018; Kayser et al., 2015; Wittmann et al., 2008). Genetic profiling studies have identified that variations in the catechol-o-methyltransferase (*COMT*) gene that increase available prefrontal dopamine are associated specifically with increases in uncertainty-driven, directed exploration (Frank et al., 2009). Novelty alone activates the nigrostriatal dopamine pathway (Bunzeck & Düzel, 2006) and this activation seems to drive novelty-driven choice behavior (Wittmann et al., 2008) characteristic of directed exploration. This is consistent with work showing that increasing available dopamine via systemic blockade of dopamine active transporter (DAT) increases the initial value monkeys assign to novel options and leads them to be more novelty-prone in their decision-making (Costa et al., 2014). However, seemingly in contrast to the findings of Wittmann et al. (2008), it has been shown that a variation in the DARPP-32 gene that increases striatal dopamine levels was associated with a decrease in directed, uncertainty-driven exploration (Gershman & Tzovaras, 2018). Here, we find that systemic increases in dopamine via L-dopa administration have no effect on tonic, directed exploration, in contrast with the findings of Chakroun et al. (2020). One difficulty in connecting our findings to the work laid out here is the gap between systemic and regional manipulations. However, given some of the bidirectional effects of dopamine on directed exploration when examined in the prefrontal cortex (Frank et al., 2009) versus in the striatum (Gershman & Tzovaras, 2018), our null finding could represent the blurring or nullifying of these competing regional effects.

With respect to random exploration, studies examining the role of dopamine have equally mixed results. For example, while some studies have shown that decreasing tonic dopamine increases random exploration (Cinotti et al., 2019), others have shown that increasing tonic dopamine has a similar effect on behavior (Beeler et al., 2012). When examining striatal dopamine, one theory is that D1-mediated signaling in the striatum regulate the randomness of action selection during explore-exploit decision-making (Humphries et al., 2012). Specifically, Humphries et al. (2012) propose that higher D1 receptor activation in the striatum leads to less random exploration, which tracks with evidence from genetic studies associating increased levels of DARPP-32 in the striatum with reductions in random exploration (Gershman & Tzovaras, 2018). It is also important to note that activation of D2 receptors in both the striatum and prefrontal cortex is often associated with increased behavioral flexibility (Barker et al., 2013; Beeler et al., 2014; Nelson & Killcross, 2013; Klanker et al., 2013), and thus likely plays a role in the willingness to explore novel options. However, in the work of Chakroun et al., (2020) systemic administration of the D2 antagonist haloperidol was found to have no effect on random or directed exploration.

With respect to tonic exploration, Ebitz et al. (2019) found that chronic cocaine self-administration decreased tonic exploratory noise. This caused animals to be less flexible both when rules changed (perseverating more) and when the environment was stable (lapsing less), but the slope of the line of best fit describing the relationship between the two types of errors was unaffected. First, this effect supports the notion that perseverative errors and lapses share an underlying exploratory cause because a single perturbation shifted them both along the same axis that they originally co-varied on (Ebitz et al., 2019). Second, the effect of cocaine on tonic exploration lends support to theories of increased tonic dopamine driving directed exploration. In contrast, our work here shows no effect of increasing tonic dopamine on tonic exploration.

It is possible that this discrepancy is due to the different mechanisms by which cocaine and L-dopa increase tonic dopamine levels. L-dopa, the dopamine precursor, acts mainly through providing a phasic dopaminergic stimulation as new dopamine is produced intracellularly and released (Robinson et al., 2005; Poletti & Bonuccelli, 2012). Cocaine, on the other hand, blocks DAT and thus increases tonic levels of extracellular dopamine (Verma, 2015). In addition, cocaine has far-reaching impacts on different aspects of phasic dopamine signaling. For example, cocaine has been shown to increase the amount of striatal dopamine produced by the excitation of dopaminergic cells in that region (Venton et al., 2006). Further, cocaine intake for as little as two weeks has been shown to prolong the activation of D1-receptors (Buchta & Riegel, 2015; Park et al., 2013) while decreasing the availability of D2 receptors (Volkow et al., 1993), shifting the balance between excitatory and inhibitory signaling. Thus, it is likely that the effects of cocaine Ebitz et al. (2019) observe are tied to one (or many) of these receptor-specific signaling changes not induced by short-term use of L-dopa. It is equally possibly that they could be due to the known effects of cocaine on noradrenergic (Beveridge et al., 2005; Burchett & Bannon, 1997; Macey et al., 2003) and cholinergic (Gifford & Johnson, 1992; Hurd et al., 1990) signaling mechanisms that we did not examine in our study.

An important caveat for our null results with respect to dopamine in the current study is that dose levels were chosen based off of those shown previously to have effects in the literature (Cools et al., 2007; Costa et al., 2015; Turchi et al., 2010). However, given the mixed results described above, as well as the differences in pre- versus post- synaptic effects for L-dopa and haloperidol, it is unclear whether different doses of these drugs would have produced different behavioral effects. For example, the effects of L-dopa on reversal learning performance have been shown to depend on baseline dopamine tone (Cools et al., 2007). Similarly, the effects of haloperidol blocking striatal D2 receptors on reversal learning performance have been shown to follow a U-shaped curve where very low and very high doses seem to impair performance, while moderate doses enhance it (Horst et al., 2019). Therefore, without completing a dose-response analysis for the effects of these two drugs or accounting for potential region-specific effects, we cannot prematurely conclude that they truly have no effect on exploratory behavior in the task.

Dopamine itself is also precursor for norepinephrine, a neuromodulator that has been implicated in regulating exploration (Yu and Dayan, 2005; Aston-Jones & Cohen. 2005). Adaptive gain theory, for example, posits that the locus coeruleus (LC) and norepinephrine system are crucial for regulating the balance between exploratory and exploitative states (Aston-Jones & Cohen, 2005). In this framework, the LC's two 'modes' map onto exploratory states. In its phasic mode, LC activity favors learning and exploitation by releasing norepinephrine when some task-relevant event occurs. In its tonic mode, LC activity has increased noise, favoring disengagement from the current behavior and pursuit of others (Aston-Jones & Cohen, 2005). Work using designer receptors exclusively activated by designer drugs (DREADDs) expressed in the LC has supported a causal relationship between increased tonic LC activity and disengagement from ongoing behavior (Kane et al., 2017). However, pharmacological studies more directly examining the explore-exploit dilemma have shown that increases in tonic norepinephrine lead to decreases in random exploration (Warren et al. 2017) or have no effect at all (Jepma et al., 2010). Many of these studies utilize systemic drugs administered in humans, and it is clear from the mixed results that more pathway- or region- specific studies are needed to parse the role that norepinephrine plays in exploration, broadly. Moreover, because none of these studies looked at tonic exploration, its relation to the LC and noradrenergic systems is unknown.

Neural circuity supporting tonic exploration

Previous work has focused on the role that various cortical regions play in exploratory behavior. Activity in prefrontal regions has been linked to evidence accumulation and evolving value representations in decision-making tasks, both of which are critical in signaling when to explore. The medial orbitofrontal and adjacent medial prefrontal cortices, known together as the ventromedial prefrontal cortex (vmPFC), have consistently been implicated in tracking evidence accumulation (Blanchard & Garshman, 2018; Vaidya & Badre, 2020) and representing decision values (Chib et al., 2009; Hare et al., 2008; Kable & Glimcher, 2007; Rudebeck et al., 2017; Smith et al., 2010). In decision-making tasks, patients with lesions of the vmPFC opt for choices yielding high immediate rewards without regard for future losses (Bechara et al., 2000), characteristic of a more exploitative state. Further, functional magnetic resonance imaging of participants with an intact vmPFC has shown that this region's activity is significantly correlated with decision values during exploitative choices (Daw et al., 2006). It has been proposed that, in line with these roles in evidence accumulation and value representation, the vmPFC encodes the reliability of a current action plan relative to expected action outcomes in such a way that signals the necessity for either exploration or exploitation (Domenech et al., 2020). Taken together, these findings point to an important role for the vmPFC in exploration that is directed at maximizing reward. Since we find here that tonic exploration is reward maximizing, it is likely that vmPFC engagement is critical to the execution of tonic exploration.

Another region consistently implicated in the switch from exploitation to exploration is the most rostral subdivision of the prefrontal cortex, commonly known as the frontopolar cortex (FPC). Because this region is important in high-level behavioral control and mediation between multiple goals (Miller et al., 2001; Ramnani et al., 2004; Koechlin and Hyafil, 2007; Boorman et al., 2009), it is an intuitive candidate for implementing exploratory states. Elevated FPC activity has been directly implicated in driving exploratory choices (Daw et al., 2006; Beharelle et al., 2015; Zajkowski et al., 2017) and, importantly, FPC activity has been shown to track trial-by-trial changes in relative uncertainty that informs exploration (Badre et al., 2012). While our measures of relative uncertainty, expressed as the posterior probabilities of reversal and of reward, are computationally different from those in Badre and colleagues' (2012) work, the joint implications of the FPC in exploratory states and in uncertainty tracking suggests this region is also involved in tonic exploration.

The dorsolateral prefrontal cortex (dlPFC) is commonly implicated in behavioral flexibility, and thus may also be important for tonic exploration. Beginning with early work showing that patients with dlPFC damage showed deficits in behavioral flexibility during a set-shifting task (Milner, 1963), the region has been implicated broadly in encoding information about environmental states and the transitions between those states (Genovesio et al., 2006; Rushworth & Behrens, 2008, Watanabe & Sakagami, 2007). For example, when human participants are making their way through a maze, their level of uncertainty about their progress through that maze correlates with activity in the dlPFC (Yoshida & Ishii, 2006). Similarly, in macaques, dlPFC activity has been associated with particular routes through mazes when animals become more certain about the most rewarding route to take (Averbeck et al., 2006).

Interestingly, Bartolo & Averbeck (2020) use the same two-armed bandit task as we do here *and* the same Bayesian model characterizing choice behavior to examine the role of the dlPFC in state switches during reversal learning. In their study, Bartolo & Averbeck (2020) recorded the activity of neural populations in both the left and right dlPFC while monkeys performed the two-armed bandit reversal learning task. They found that the activity of a group of neurons in this region corresponded with the posterior probability of a reversal – the same posterior probability from our behavioral choice model (M = 2) that we use to predict lapses. Looking more closely at the neural signals associated with this posterior, Bartolo & Averbeck (2020) found that the dlPFC signal specifically emerged in the trial before the animal switched its choice behavior, directly implicating the region in some state-switching process that supports reversal in the two-armed bandit task. To investigate whether this region also supports tonic exploration in the task, future work could test whether the dlPFC activity, in addition to corresponding with the posterior probability of a reversal before the switch in choice behavior occurs, also corresponds with that same posterior before lapses occur. For example, given how phasic exploration (i.e. how animals reverse their behavior) is negatively correlated with lapses in behavior, neural activity in the DLPFC that encodes the switch may be negatively correlated with activity in the same region during lapses if this region is involved in tonic exploration.

While it is often assumed that cortical areas are the primary drivers of higher order behaviors such as deciding whether to explore, recent work has also highlighted the contributions of subcortical areas to exploration. For example, both the amygdala (Averbeck and Costa, 2017; Belova et al., 2008; Morrison and Salzman, 2010; Paton et al., 2006) and the ventral striatum

(Cai et al., 2011; Jocham et al., 2011; Shidara et al., 1998; Simmons et al., 2007; Strait et al., 2015) represent value across a range of decision-making tasks. In a study by Costa et al. (2016) using the same two-armed bandit reversal learning task we used here, animals performed the task after receiving lesions of either the amygdala or the ventral striatum (VS). Costa et al. (2016) found that animals with lesions of the VS and animals with lesions of the amygdala chose the high value option less consistently compared to control animals in all three of the probabilistic reward schedules (80/20, 70/30, 60/40). While this decreased choice consistency may indicate that the animals simply had more randomness in their choice behavior, it may also reflect increased tonic exploration when the amygdala or ventral striatum is lesioned.

The effects of striatal lesions that Costa et al. (2016) observe are consistent with the theory that reduced D1-mediated signaling in the striatum increases the randomness of action selection during explore-exploit decision-making (Humphries et al., 2012). It is possible, then, that decreased choice consistency in VS-lesioned animals represents an increase in tonic exploration stemming from the reduced influence of D1 projections to cortical areas. For example, given the known connectivity between ventral areas of the striatum and the vmPFC (Haber, 2016) as well as the role of the vmPFC in representing value during exploitative choice (Daw et al., 2006), reduced D1-mediated activation of vmPFC by the VS may reduce animals' exploitative tendencies and in turn generate tonic exploratory noise in behavior. Similarly, work in rodents has shown that D1 receptors are primarily expressed in the basolateral nucleus of the amygdala (Abraham et al., 2014; Muly et al., 2009; Weiner et al. 1991), which sends projections to the medial prefrontal cortex in rodents (Hoover & Vertes, 2007) and in primates (Kim et al. 2011). Thus, a similar mechanism could be at play in the amygdala-lesioned animals whose choice consistency was lower, where reduced D1 input to medial prefrontal cortices increases tonic exploratory noise in behavior. However, without applying similar behavioral analyses as we do here to rule out nuisance causes of this randomness in behavior, it is unclear whether or not the effects of subcortical lesions on choice consistency represent an effect of cortico-striatal/amygdalo-cortical dopaminergic projections on tonic exploration. To test whether the noisy decision-making in the lesion groups' behavior is due to tonic exploration, future work could use the same methods as we did here (i.e. correlating phasic exploration with lapses during the reversal phase of the task) and observe the relationship. First, these analyses would need to establish a negative correlation in control animals between the two measures to establish that tonic exploration is being used in the task. Second, these analyses could examine if and how lesions of the VS or amygdala affect this relationship correlation in order to determine possible effects of either region on tonic exploration.

Summary and Conclusions

In summary, lapses in decision-making during the reversal phase of a two-armed bandit may serve as a behavioral metric for tonic, directed exploration. Adding to the existing literature, we show that tonic exploration is a viable strategy in probabilistic reward environments and that its implementation is not dependent on the overall levels of expected uncertainty in the environment. Importantly, we also find that L-dopa and haloperidol have no effect on the use of tonic exploration. Additional studies are needed to validate the use of tonic exploration across experimental paradigms and across species, as well as to elucidate the mechanisms that support this kind of behavior.

Given the importance of uncertainty in driving exploration, future studies should assess if and how tonic exploration is employed with respect to unexpected uncertainty, volatility, and unpredictability in stimulus-outcome or action-outcome relationships. Additionally, our lack of dopamine effects here stand in contrast to other work that has identified dopamine as a critical modulator of the explore-exploit tradeoff. Future studies must take advantage of more targeted (i.e. regional, chemogenetic) manipulations to parse out the role of dopamine (or norepinephrine) in regulating tonic exploration. Lastly, future studies should leverage imaging techniques as well as manipulating region- and pathway-specific activity during tasks where tonic exploration is used to better understand the neuroanatomy that supports tonic exploration.

In conclusion, the work outlined in this thesis highlights the importance of incorporating measures of tonic exploration into solutions of the explore-exploit dilemma. By showing that tonic exploration is being employed to reduce uncertainty in complex learning environments, we demonstrate the potential relevance of tonic exploratory noise to current theories of exploratory decision-making.

References

Abraham, A. D., Neve, K. A., & Lattal, K. M. (2014). Dopamine and extinction: a convergence of theory with fear and reward circuitry. *Neurobiology of learning and memory, 108*, 65–77. https://doi.org/10.1016/j.nlm.2013.11.007

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacol*, *42,* 1931-1939. https://doi.org/10.1038/npp.2017.108

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Averbeck, B. B. (2015). Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLOS Computational Biology*, *11*(3), e1004164. https://doi.org/10.1371/journal.pcbi.1004164

Averbeck, B. B., & Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience*, *20*(4), 505–512. https://doi.org/10.1038/nn.4506

Averbeck, B.B., Sohn, J.W. & Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience (9),* 276–282. https://doi.org/10.1038/nn1634

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron*, *73*(3), 595–607. https://doi.org/10.1016/j.neuron.2011.12.025

Barack, D. L., & Gold, J. I. (2016). Temporal trade-offs in psychophysics. *Current Opinion in Neurobiology*, *37*, 121–125. https://doi.org/10.1016/j.conb.2016.01.015

Bartolo, R., & Averbeck, B. B. (2020). Prefrontal Cortex Predicts State Switches during Reversal Learning. *Neuron*, *106*(6), 1044-1054.e4. https://doi.org/10.1016/j.neuron.2020.03.024

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. Journal of the *Experimental Analysis of Behavior, 32(2),* 269–281. https://doi:10.1901/jeab.1979.32-269

Baum W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the experimental analysis of behavior*, *22*(1), 231–242. https://doi.org/10.1901/jeab.1974.22-231

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, *123*(11), 2189–2202. https://doi.org/10.1093/brain/123.11.2189

Beeler, J. A., Cools, R., Luciana, M., Ostlund, S. B., & Petzinger, G. (2014). A kinder, gentler dopamine… highlighting dopamine's role in behavioral flexibility. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00004

Beeler, J. A., Daw, N. D., Frazier, C. R. M., & Zhuang, X. (2010a). Tonic Dopamine Modulates Exploitation of Reward Learning. *Frontiers in Behavioral Neuroscience*, *4*. https://doi.org/10.3389/fnbeh.2010.00170

Beharelle, A. R., Polanía, R., Hare, T. A., & Ruff, C. C. (2015). Transcranial Stimulation over Frontopolar Cortex Elucidates the Choice Attributes and Neural Mechanisms Used to Resolve Exploration–Exploitation Trade-Offs. *The Journal of Neuroscience*, *35*(43), 14544–14556. https://doi.org/10.1523/JNEUROSCI.2322-15.2015

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-Moment Tracking of State Value in the Amygdala. *The Journal of Neuroscience*, *28*(40), 10023–10030. https://doi.org/10.1523/JNEUROSCI.1400-08.2008

Beveridge, T. J. R., Smith, H. R., Nader, M. A., & Porrino, L. J. (2005). Effects of chronic cocaine self-administration on norepinephrine transporters in the nonhuman primate brain. *Psychopharmacology*, *180*(4), 781–788. https://doi.org/10.1007/s00213-005-2162-1

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(1), 117–126. https://doi.org/10.3758/s13415-017-0556-2

Bland, A. R., & Schaefer, A. (2012). Different Varieties of Uncertainty in Human Decision-Making. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00085

Bonuccelli, U., & Pavese, N. (2006). Dopamine agonists in the treatment of Parkinson's disease. *Expert Review of Neurotherapeutics*, *6*(1), 81–89. https://doi.org/10.1586/14737175.6.1.81

Buchta, W. C., & Riegel, A. C. (2015). Chronic cocaine disrupts mesocortical learning mechanisms. *Brain Research*, *1628*(0 0), 88–103. https://doi.org/10.1016/j.brainres.2015.02.003

Bunzeck, N., & Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, *51*(3), 369–379. https://doi.org/10.1016/j.neuron.2006.06.021

Burchett, S. A., & Bannon, M. J. (1997). Serotonin, dopamine and norepinephrine transporter mRNAs: Heterogeneity of distribution and response to `binge' cocaine administration. *Molecular Brain Research*, *49*(1), 95–102. https://doi.org/10.1016/S0169-328X(97)00131-9

Cai, X., Kim, S., & Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron*, *69*(1), 170–182. https://doi.org/10.1016/j.neuron.2010.11.041

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *ELife*, *9*, e51260. https://doi.org/10.7554/eLife.51260

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, *9*(1), 6770. https://doi.org/10.1038/s41598-019-43245-z

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage*, *35*(2), 968–978. https://doi.org/10.1016/j.neuroimage.2006.11.056

Collins, A. G. E., & Frank, M. J. (2016). Surprise! Dopamine signals mix action, value and error. *Nature Neuroscience*, *19*(1), 3–5. https://doi.org/10.1038/nn.4207

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective & Behavioral Neuroscience*, *15*(4), 837–853. https://doi.org/10.3758/s13415-015-0350-y

Cools, R., Altamirano, L., & D'Esposito, M. (2006). Reversal learning in Parkinson's disease depends on medication status and outcome valence. *Neuropsychologia*, *44*(10), 1663–1673. https://doi.org/10.1016/j.neuropsychologia.2006.03.030

Cools, R., Lewis, S. J. G., Clark, L., Barker, R. A., & Robbins, T. W. (2007). L -DOPA Disrupts Activity in the Nucleus Accumbens during Reversal Learning in Parkinson's Disease. *Neuropsychopharmacology*, *32*(1), 180–189. https://doi.org/10.1038/sj.npp.1301153

Costa, V. D., & Averbeck, B. B. (2020). Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore–Exploit Tradeoffs. *The Journal of Neuroscience*, *40*(12), 2553–2561. https://doi.org/10.1523/JNEUROSCI.2355-19.2020

Costa, V. D., Mitz, A. R., & Averbeck, B. B. (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. *Neuron*, *103*(3), 533-545.e5. https://doi.org/10.1016/j.neuron.2019.05.017

Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. (2014). Dopamine modulates novelty seeking behavior during decision making. *Behavioral Neuroscience*, *128*(5), 556–566. https://doi.org/10.1037/a0037128

Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. (2015). Reversal learning and dopamine: A bayesian perspective. *The Journal of Neuroscience*, *35*(6), 2407–2416. https://doi.org/10.1523/JNEUROSCI.1989-14.2015

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

de Villiers, P. A., & Herrnstein, R. J. (1976). Toward a law of response strength. *Psychological Bulletin, 83*(6), 1131–1153. https://doi.org/10.1037/0033-2909.83.6.1131

Domenech, P., Rheims, S., & Koechlin, E. (2020). Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science*, *369*(6507). https://doi.org/10.1126/science.abb0184

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. https://doi.org/10.1038/nn2077

Ebitz, R. B., Albarran, E., & Moore, T. (2018). Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex. *Neuron*, *97*(2), 450-461.e9. https://doi.org/10.1016/j.neuron.2017.12.007

Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W., & Hayden, B. Y. (2019). Tonic exploration governs both flexibility and lapses. *PLOS Computational Biology*, *15*(11), e1007475. https://doi.org/10.1371/journal.pcbi.1007475

Everitt, B. J., & Robbins, T. W. (2016). Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual Review of Psychology*, *67*(1), 23–50. https://doi.org/10.1146/annurev-psych-122414-033457

Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2021). The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific Reports*, *11*(1), 3077. https://doi.org/10.1038/s41598-021-82530-8

Finetti, B. de. (2017). *Theory of Probability: A critical introductory treatment*. John Wiley & Sons.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, *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. M., & Akil, H. (2011). A selective role for dopamine in stimulus–reward learning. *Nature*, *469*(7328), 53–57. https://doi.org/10.1038/nature09588

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). The neurogenetics of exploration and exploitation: Prefrontal and striatal dopaminergic components. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Frank, M.J., O'Reilly, R.C. (2006). A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. *Behav Neurosci 120*, 497–517. https://doi.apa.org/doi/10.1037/0735-7044.120.3.497

Genovesio, A., Brasted, P. J., & Wise, S. P. (2006). Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. *The Journal of Neuroscience, 26(27),* 7305–7316. https://doi.org/10.1523/JNEUROSCI.0699-06.2006

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J., & Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, *120*, 97–104. https://doi.org/10.1016/j.neuropsychologia.2018.10.009

Gifford, A. N., & Johnson, K. M. (1992). Effect of chronic cocaine treatment on D2 receptors regulating the release of dopamine and acetylcholine in the nucleus accumbens and striatum. *Pharmacology Biochemistry and Behavior*, *41*(4), 841–846. https://doi.org/10.1016/0091-3057(92)90236-9

Gittins, J. C. (1979). Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164. https://doi.org/10.1111/j.2517-6161.1979.tb01068.x

Haber S. N. (2016). Corticostriatal circuitry. *Dialogues in clinical neuroscience, 18(1),* 7–21. https://doi.org/10.31887/DCNS.2016.18.1/shaber

Haluk, D. M., & Floresco, S. B. (2009). Ventral Striatal Dopamine Modulation of Different Forms of Behavioral Flexibility. *Neuropsychopharmacology*, *34*(8), 2041–2052. https://doi.org/10.1038/npp.2009.21

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

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *Journal of Neuroscience*, *28*(22), 5623–5630. https://doi.org/10.1523/JNEUROSCI.1309-08.2008

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, *4*(3), 267–272. https://doi.org/10.1901/jeab.1961.4-267

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and Matching on Concurrent Ratio Schedules1. *Journal of the Experimental Analysis of Behavior*, *24*(1), 107–116. https://doi.org/10.1901/jeab.1975.24-107

Horst, N.K., Jupp, B., Roberts, A.C., & Robbins, T. W. (2019). D2 receptors and cognitive flexibility in marmosets: tri-phasic dose–response effects of intra-striatal quinpirole on serial reversal performance. *Neuropsychopharmacol 44*, 564–571. https://doi.org/10.1038/s41386-018-0272-9

Hoover, W. B., & Vertes, R. P. (2007). Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. Brain structure & function, 212(2), 149–179. https://doi.org/10.1007/s00429-007-0150-4

Huang, X., Gu, H. H., & Zhan, C.-G. (2009). Mechanism for Cocaine Blocking the Transport of Dopamine: Insights from Molecular Modeling and Dynamics Simulations. *The Journal of Physical Chemistry. B*, *113*(45), 15057–15066. https://doi.org/10.1021/jp900963n

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00009

Hurd, Y. L., Weiss, F., Koob, G., & Ungerstedt, U. (1990). The influence of cocaine self-administration on in vivo dopamine and acetylcholine neurotransmission in rat caudate-putamen. *Neuroscience Letters*, *109*(1), 227–233. https://doi.org/10.1016/0304-3940(90)90568-T

Iigaya, K., Ahmadian, Y., Sugrue, L. P., Corrado, G. S., Loewenstein, Y., Newsome, W. T., & Fusi, S. (2019). Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. *Nature communications*, *10*(1), 1466. https://doi.org/10.1038/s41467-019-09388-3

Jenni, N. L., Larkin, J. D., & Floresco, S. B. (2017). Prefrontal Dopamine D1 and D2 Receptors Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and Amygdalar Circuits. *Journal of Neuroscience*, *37*(26), 6200–6213. https://doi.org/10.1523/JNEUROSCI.0030-17.2017

Jepma, M., Te Beek, E. T., Wagenmakers, E.-J., Van Gerven, J. M. A., & Nieuwenhuis, S. (2010). The role of the noradrenergic system in the exploration-exploitation trade-off: A pharmacological study. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00170

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices. *Journal of Neuroscience*, *31*(5), 1606–1613. https://doi.org/10.1523/JNEUROSCI.3904-10.2011

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633. https://doi.org/10.1038/nn2007

Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective & Behavioral Neuroscience*, *17*(6), 1073–1083. https://doi.org/10.3758/s13415-017-0531-y

Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, Locus of Control, and the Exploration-Exploitation Tradeoff. *Neuropsychopharmacology*, *40*(2), 454–462. https://doi.org/10.1038/npp.2014.193

Kim, M. J., Loucks, R. A., Palmer, A. L., Brown, A. C., Solomon, K. M., Marchante, A. N., & Whalen, P. J. (2011). The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. *Behavioural brain research, 223(2),* 403–410. https://doi.org/10.1016/j.bbr.2011.04.025

Klanker, M., Feenstra, M., & Denys, D. (2013). Dopaminergic control of cognitive flexibility in humans and animals. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00201

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, *275*(5675), 27–31. https://doi.org/10.1038/275027a0

Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S.-C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17951–17956. https://doi.org/10.1073/pnas.0905191106

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, *84*(3), 555–579. https://doi.org/10.1901/jeab.2005.110-04

Lee, E., Seo, M., Dal Monte, O., & Averbeck, B. B. (2015). Injection of a dopamine type 2 receptor antagonist into the dorsal striatum disrupts choices driven by previous outcomes, but not perceptual inference. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(16), 6298–6306. https://doi.org/10.1523/JNEUROSCI.4561-14.2015

Macey, D. J., Smith, H. R., Nader, M. A., & Porrino, L. J. (2003). Chronic Cocaine Self-Administration Upregulates the Norepinephrine Transporter and Alters Functional Activity in the Bed Nucleus of the Stria Terminalis of the Rhesus Monkey. *The Journal of Neuroscience*, *23*(1), 12. https://doi.org/10.1523/JNEUROSCI.23-01-00012.2003

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, *84*(4), 870–881. https://doi.org/10.1016/j.neuron.2014.10.013

McNamee, D., Rangel, A., & O'Doherty, J. P. (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nature Neuroscience*, *16*(4), 479–485. https://doi.org/10.1038/nn.3337

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Arch Neurol., 9(1),* 90–100. https://10.1001/archneur.1963.00460070100010

Moore, T. L., Killiany, R. J., Herndon, J. G., Rosene, D. L., & Moss, M. B. (2005). A non-human primate test of abstraction and set shifting: An automated adaptation of the Wisconsin Card Sorting Test. *Journal of Neuroscience Methods*, *146*(2), 165–173. https://doi.org/10.1016/j.jneumeth.2005.02.005

Morris, C. J. (1987). The operant conditioning of response variability: Free-operant versus discrete-response procedures. *Journal of the Experimental Analysis of Behavior*, *47*(3), 273–277. https://doi.org/10.1901/jeab.1987.47-273

Morris, L. S., Kundu, P., Baek, K., Irvine, M. A., Mechelmans, D. J., Wood, J., Harrison, N. A., Robbins, T. W., Bullmore, E. T., & Voon, V. (2016). Jumping the Gun: Mapping Neural Correlates of Waiting Impulsivity and Relevance Across Alcohol Misuse. *Biological Psychiatry*, *79*(6), 499–507. https://doi.org/10.1016/j.biopsych.2015.06.009

Morrison, S. E., & Salzman, C. D. (2010). Re-valuing the amygdala. *Current Opinion in Neurobiology*, *20*(2), 221–230. https://doi.org/10.1016/j.conb.2010.02.007

Muly, E. C., Senyuz, M., Khan, Z. U., Guo, J. D., Hazra, R., & Rainnie, D. G. (2009). Distribution of D1 and D5 dopamine receptors in the primate and rat basolateral amygdala. *Brain structure & function*, *213*(4-5), 375–393. https://doi.org/10.1007/s00429-009-0214-8

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

Nelson, A. J. D., & Killcross, S. (2013). Accelerated habit formation following amphetamine exposure is reversed by D1, but enhanced by D2, receptor antagonists. *Frontiers in Neuroscience*, *7*. https://doi.org/10.3389/fnins.2013.00076

O’Farrell, S., Sanchirico, J. N., Spiegel, O., Depalle, M., Haynie, A. C., Murawski, S. A., Perruso, L., & Strelcheck, A. (2019). Disturbance modifies payoffs in the explore-exploit trade-off. *Nature Communications*, *10*(1), 3363. https://doi.org/10.1038/s41467-019-11106-y

Park, K., Volkow, N. D., Pan, Y., & Du, C. (2013). Chronic Cocaine Dampens Dopamine Signaling during Cocaine Intoxication and Unbalances D1 over D2 Receptor Signaling. *The Journal of Neuroscience*, *33*(40), 15827–15836. https://doi.org/10.1523/JNEUROSCI.1935-13.2013

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865–870. https://doi.org/10.1038/nature04490

Paulus, M. P., Feinstein, J. S., Simmons, A., & Stein, M. B. (2004). Anterior cingulate activation in high trait anxious subjects is related to altered error processing during decision making. *Biological Psychiatry*, *55*(12), 1179–1187. https://doi.org/10.1016/j.biopsych.2004.02.023

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *PLOS Computational Biology*, *7*(1), e1001048. https://doi.org/10.1371/journal.pcbi.1001048

Pierce, W. D., & Epling, W. F. (1983). Choice, matching, and human behavior: A review of the literature. *The Behavior analyst*, *6*(1), 57–76. https://doi.org/10.1007/BF03391874

Piray, P., & Daw, N. D. (2020a). A simple model for learning in volatile environments. *PLOS Computational Biology*, *16*(7), e1007963. https://doi.org/10.1371/journal.pcbi.1007963

Piray, P., & Daw, N. D. (2020b). Unpredictability vs. Volatility and the control of learning. *BioRxiv*, 2020.10.05.327007. https://doi.org/10.1101/2020.10.05.327007

Poletti, M., & Bonuccelli, U. (2013). Acute and chronic cognitive effects of levodopa and dopamine agonists on patients with Parkinson's disease: A review. *Therapeutic Advances in Psychopharmacology*, *3*(2), 101–113. https://doi.org/10.1177/2045125312470130

Polezzi, D., Lotto, L., Daum, I., Sartori, G., & Rumiati, R. (2008). Predicting outcomes of decisions in the brain. *Behavioural Brain Research*, *187*(1), 116–122. https://doi.org/10.1016/j.bbr.2007.09.001

Rahman, S., & Marwaha, R. (2021). Haloperidol. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK560892/

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, *5*(3), 184–194. https://doi.org/10.1038/nrn1343

Ridley, R. M., Haystead, T. A. J., & Baker, H. F. (1981). An analysis of visual object reversal learning in the marmoset after amphetamine and haloperidol. *Pharmacology Biochemistry and Behavior*, *14*(3), 345–351. https://doi.org/10.1016/0091-3057(81)90401-9

Robinson, S., Sandstrom, S. M., Denenberg, V. H., & Palmiter, R. D. (2005). Distinguishing Whether Dopamine Regulates Liking, Wanting, and/or Learning About Rewards. *Behavioral Neuroscience*, *119*(1), 5–15. https://doi.org/10.1037/0735-7044.119.1.5

Rudebeck, P. H., Saunders, R. C., Lundgren, D. A., & Murray, E. A. (2017). Specialized Representations of Value in the Orbital and Ventrolateral Prefrontal Cortex: Desirability versus Availability of Outcomes. *Neuron*, *95*(5), 1208-1220.e5. https://doi.org/10.1016/j.neuron.2017.07.042

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397. https://doi.org/10.1038/nn2066

Salamone, J. D., Correa, M., Mingote, S. M., & Weber, S. M. (2005). Beyond the reward hypothesis: Alternative functions of nucleus accumbens dopamine. *Current Opinion in Pharmacology*, *5*(1), 34–41. https://doi.org/10.1016/j.coph.2004.09.004

Scholl, J., & Klein-Flügge, M. (2018). Understanding psychiatric disorder by capturing ecologically relevant features of learning and decision-making. *Behavioural Brain Research*, *355*, 56–75. https://doi.org/10.1016/j.bbr.2017.09.050

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Shidara, M., Aigner, T. G., & Richmond, B. J. (1998). Neuronal Signals in the Monkey Ventral Striatum Related to Progress through a Predictable Series of Trials. *The Journal of Neuroscience*, *18*(7), 2613–2625. https://doi.org/10.1523/JNEUROSCI.18-07-02613.1998

Shohamy, D., Myers, C. E., Geghman, K. D., Sage, J., & Gluck, M. A. (2006). L-dopa impairs learning, but spares generalization, in Parkinson’s disease. *Neuropsychologia*, *44*(5), 774–784. https://doi.org/10.1016/j.neuropsychologia.2005.07.013

Simmons, J. M., Ravel, S., Shidara, M., & Richmond, B. J. (2007). A comparison of reward-contingent neuronal activity in monkey orbitofrontal cortex and ventral striatum: Guiding actions toward rewards. *Annals of the New York Academy of Sciences*, *1121*, 376–394. https://doi.org/10.1196/annals.1401.028

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J. A., Pitchford, J. W., James, A., Ahmed, M. Z., Brierley, A. S., Hindell, M. A., Morritt, D., Musyl, M. K., Righton, D., Shepard, E. L. C., Wearmouth, V. J., Wilson, R. P., Witt, M. J., & Metcalfe, J. D. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098–1102. https://doi.org/10.1038/nature06518

Smith, D. V., Hayden, B. Y., Truong, T.-K., Song, A. W., Platt, M. L., & Huettel, S. A. (2010). Distinct Value Signals in Anterior and Posterior Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *30*(7), 2490–2495. https://doi.org/10.1523/JNEUROSCI.3319-09.2010

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews. Neuroscience*, *20*(10), 635–644. https://doi.org/10.1038/s41583-019-0180-y

Strait, C. E., Sleezer, B. J., & Hayden, B. Y. (2015). Signatures of Value Comparison in Ventral Striatum Neurons. *PLOS Biology*, *13*(6), e1002173. https://doi.org/10.1371/journal.pbio.1002173

Strauss, G. P., Frank, M. J., Waltz, J. A., Kasanova, Z., Herbener, E. S., & Gold, J. M. (2011). Deficits in Positive Reinforcement Learning and Uncertainty-Driven Exploration are Associated with Distinct Aspects of Negative Symptoms in Schizophrenia. *Biological Psychiatry*, *69*(5), 424–431. https://doi.org/10.1016/j.biopsych.2010.10.015

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement Learning: An Introduction.* MIT Press: Cambridge, MA, USA.

Trudel, N., Scholl, J., Klein-Flügge, M. C., Fouragnan, E., Tankelevitch, L., Wittmann, M. K., & Rushworth, M. F. S. (2021). Polarity of uncertainty representation during exploration and exploitation in ventromedial prefrontal cortex. *Nature Human Behaviour*, *5*(1), 83–98. https://doi.org/10.1038/s41562-020-0929-3

Trepka, E., Spitmaan, M., Bari, B. A., Costa, V. D., Cohen, J. Y., & Soltani, A. (2021). Novel entropy-based metrics for predicting choice behavior based on local response to reward. *BioRxiv*, 2021.05.20.445009. https://doi.org/10.1101/2021.05.20.445009

Vaidya, A. R., & Badre, D. (2020). Neural systems for memory-based value judgment and decision-making. *BioRxiv*, 712661. https://doi.org/10.1101/712661

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. (2014). A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

van Holstein, M., Aarts, E., van der Schaaf, M. E., Geurts, D. E. M., Verkes, R. J., Franke, B., van Schouwenburg, M. R., & Cools, R. (2011). Human cognitive flexibility depends on dopamine D2 receptor signaling. *Psychopharmacology*, *218*(3), 567–578. https://doi.org/10.1007/s00213-011-2340-2

Venton, B. J., Seipel, A. T., Phillips, P. E. M., Wetsel, W. C., Gitler, D., Greengard, P., Augustine, G. J., & Wightman, R. M. (2006). Cocaine Increases Dopamine Release by Mobilization of a Synapsin-Dependent Reserve Pool. *Journal of Neuroscience*, *26*(12), 3206–3209. https://doi.org/10.1523/JNEUROSCI.4901-04.2006

Volkow, N. D., Fowler, J. S., Wang, G. J., Hitzemann, R., Logan, J., Schlyer, D. J., Dewey, S. L., & Wolf, A. P. (1993). Decreased dopamine D2 receptor availability is associated with reduced frontal metabolism in cocaine abusers. *Synapse (New York, N.Y.)*, *14*(2), 169–177. https://doi.org/10.1002/syn.890140210

Volz, K. G., Schubotz, R. I., & von Cramon, D. Y. (2003). Predicting events of varying probability: Uncertainty investigated by fMRI. *NeuroImage*, *19*(2), 271–280. https://doi.org/10.1016/S1053-8119(03)00122-8

Watanabe, M., & Sakagami, M. (2007). Integration of cognitive and motivational context information in the primate prefrontal cortex. *Cerebral cortex*, 17(Suppl 1), i101–i109. https://doi.org/10.1093/cercor/bhm067

Warren, C. M., Wilson, R. C., Wee, N. J. van der, Giltay, E. J., Noorden, M. S. van, Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PLOS ONE*, *12*(4), e0176034. https://doi.org/10.1371/journal.pone.0176034

Wearden, J. H. (1983). Undermatching and overmatching as deviations from the matching law. *Journal of the Experimental Analysis of Behavior*, *40*(3), 333–340. https://doi.org/10.1901/jeab.1983.40-333

Weiner, D. M., Levey, A. I., Sunahara, R. K., Niznik, H. B., O'Dowd, B. F., Seeman, P., & Brann, M. R. (1991). D1 and D2 dopamine receptor mRNA in rat brain. *Proceedings of the National Academy of Sciences of the United States of America, 88(5),* 1859–1863. https://doi.org/10.1073/pnas.88.5.1859

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology. General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008). Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron*, *58*(6), 967–973. https://doi.org/10.1016/j.neuron.2008.04.027

Yoshida, W. & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron(50),* 781–789. https://10.1016/j.neuron.2006.05.006

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (n.d.). A causal role for right frontopolar cortex in directed, but not random, exploration. *ELife*, *6*. https://doi.org/10.7554/eLife.27430

Zhukovsky, P., Puaud, M., Jupp, B., Sala-Bayo, J., Alsiö, J., Xia, J., Searle, L., Morris, Z., Sabir, A., Giuliano, C., Everitt, B. J., Belin, D., Robbins, T. W., & Dalley, J. W. (2019). Withdrawal from escalated cocaine self-administration impairs reversal learning by disrupting the effects of negative feedback on reward exploitation: A behavioral and computational analysis. *Neuropsychopharmacology*, *44*(13), 2163–2173. https://doi.org/10.1038/s41386-019-0381-0

Zuriff, G. E. (1970). A comparison of variable-ratio and variable-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior*, *13*(3), 369–374. https://doi.org/10.1901/jeab.1970.13-369