The Effects of Uncertainty on Behavior: Reward Distributions Modulate Dopamine Learning Signals and Decision Making

by

Kathryn M. Rothenhoefer

BA Psychology, American University, 2015

Submitted to the Graduate Faculty of the

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Kathryn M. Rothenhoefer

It was defended on

November 28, 2022

and approved by

Committee Chair: Aaron P. Batista, PhD, Bioengineering, University of Pittsburgh

Carl R. Olson, PhD, Biomedical Engineering, Carnegie Mellon University

Caroline A. Runyan, PhD, Neuroscience, University of Pittsburgh

Afonso C. Silva, PhD, Neurobiology, University of Pittsburgh

Erin L. Rich, MD, PhD, Neuroscience, Icahn School of Medicine at Mount Sinai

Dissertation Advisor: William R. Stauffer, PhD, Neurobiology, University of Pittsburgh

The Effects of Uncertainty on Behavior: Reward Distributions Modulate Dopamine Learning Signals and Decision Making

Kathryn M. Rothenhoefer

University of Pittsburgh, 2022

ABSTRACT

To make optimal decisions and get the best outcomes, human and animal decision-makers must traverse an uncertain world. Uncertainty is divided into two separate types: risk and ambiguity. Risk describes uncertainty where the outcomes and their probabilities are known, such as a fair coin flip. Ambiguity is uncertainty where there is incomplete information about the possible outcomes and their probabilities of occurring. Decision-makers often avoid choices with ambiguity, even if they are objectively better than the alternatives. Our ability to make the best decisions given the uncertainty of our options is dependent on learning from our past decisions and updating our expectations accordingly. Learning which behaviors to repeat, and which to discontinue is dependent on learning signals from midbrain dopamine neurons. Dopamine neurons compute reward prediction errors and transmit this signal to various brain regions, but mostly to the striatum, a critical mediator of reward learning. Reward predictions errors code the difference between the reward that was actually received and the reward that was expected. Expected value estimates in dopamine neuron signals incorporate the probability and size of rewards. However, it is not known whether dopamine neurons code uncertainty independent from expected value. The results of this body of work are threefold. First, we determined that higher uncertainty, when independent from the expected value of the cue, decreased learning and autonomic responses, and minimized the absolute magnitude of dopamine neuron reward prediction error responses. Second,

we determined that animals differentiate between ambiguous and risky choice options that have the same expected value. Choice behavior illustrated that ambiguity preferences vary based on expected value. This result is the first to demonstrate that decision-makers have value-dependent ambiguity preferences, like they do risk preferences. Finally, we characterized the functional and anatomical diversity of neurons in the striatum, specifically medium spiny neurons, which are a key relay station between dopamine neurons and the cortico-basal ganglia-thalamo-cortical circuits that mediate sophisticated behavior. These results together provide clear direction for future research into the effect of uncertainty on behavior and neural coding, and new opportunities for investigating circuit-specific functions that mediate these behaviors.

Acknowledgements

I would like to express my gratitude to my mentor, Bill Stauffer, for supporting me and molding me into the scientist I am today. I will always appreciate and remember our many conversations about my projects, dopamine, science, and life in general. I have learned so much.

To my family, thank you for always loving and supporting me. Mom and Dad, the gratitude I have for all the sacrifices you have made to get me on this path could never be overstated. Amy, thank you for providing support that was just a phone call away, and always being able to make me laugh when things were tough.

To my friends, thank you for always encouraging me through this process. To Luana Correia, and our cat George, thank you for always sticking by my side. To Kim Berry, Nick Chehade, and Aaron Dean, I will always cherish our friendship and camaraderie through the years here at Pitt. To Linda Amarante, I appreciate our friendship and the huge part you played in fostering my love of neuroscience as an undergrad.

To Bruno Averbeck and Vinny Costa, I truly appreciate the integral role you both played in my scientific development as a postbac, and for your continued mentorship throughout grad school.

To the members of the Stauffer Lab, you all made the work so enjoyable throughout the years, and I am so excited to see where your time in the lab takes you. To Jackie Breter, your support and friendship has been invaluable to my progress through grad school. To Andreea Bostan, our friendship and your expertise in all things neuroscience has been so greatly appreciated.

I would like to thank all the administrators in the CNUP, CNBC, and 4th Floor BST3, as well as SNARL and our veterinarians, and the custodial staff.

To my committee, thank you for your time, and the encouragement, guidance, and support you've given me through this process.

This work was funded by grants from the University of Pittsburgh Brain Institute, NIH DP2MH113095, NIMH/NIH UG3MH120094 and NIMH/NIH R01MH128669 to Bill Stauffer.

Table of Contents

List of Figures

List of Abbreviations

A1	Primary Auditory Cortex
BA	Broadmann Area
BG	Basal Ganglia
BOLD	Blood-Oxygen-Level Dependent
Cd	Caudate
CE	Certainty Equivalent
CR	Condition Response
CS	Conditioned Stimulus
dlPFC	Dorsolateral Prefrontal Cortex
DS	Dorsal Striatum
EMG	Electromyography
EU	Expected Utility
EUT	Expected Utility Theory
EV	Expected Value
FEV	Future Expected Value
fMRI	Functional Magnetic Resonance Imaging
GPe	External Segment of the Globus Pallidus
GPi	Internal Segment of the Globus Pallidus
IAL	Interaural Line
ICj	Island of Calleja

IEV	Immediate Expected Value
LC	Locus Coeruleus
LHb	Lateral Habenula
LIP	Lateral Intraparietal Cortex
lOFC	Lateral Orbitofrontal Cortex
lPFC	Lateral Prefrontal Cortex
M1	Primary Motor Cortex
MDP	Markov Decision Process
MEC	Medial Entorhinal Cortex
mPFC	Medial Prefrontal Cortex
MSN	Medium Spiny Neuron
NAc	Nucleus Accumbens
NHP	Non-Human Primate
NS	Neutral Stimulus
NUDAP	Neurochemically Unique Domains in the Accumbens and Putamen
OFC	Orbitofrontal Cortex
OT	Olfactory Tubercle
PAN	Phasically Active Neuron
PCC	Posterior Cingulate Cortex
PE	Prediction Error
PFC	Prefrontal Cortex
POMDP	Partially Observable Markov Decision Process
PSTH	Peri-Stimulus Time Histogram

Pt	Putamen
RL	Reinforcement Learning
RMTg	Rostromedial Tegmental Nucleus
ROI	Region of Interest
RPE	Reward Prediction Error
SD	Standard Deviation
SI	Selectivity Index
scRNA-seq	Single Cell RNA Sequencing
snRNA-seq	Single Nucleus RNA Sequencing
TAN	Tonically Active Neuron
TD	Temporal Difference
UR	Unconditioned Response
US	Unconditioned Stimulus
V1	Primary Visual Cortex
vmPFC	Ventromedial Prefrontal Cortex
VP	Ventral Pallidum
VS	Ventral Striatum
VTA	Ventral Tegmental Area

1.0 Introduction

Uncertainty is a pervasive aspect of decision making processes for humans and animals, and for all types of choices. In fact, very few decisions can ever be made with complete certainty of the outcome. Typically, choices are made with some uncertainty about the identity of the possible outcomes, or about the outcome likelihoods. Formalized uncertainty in decision making can be broken into two distinct forms: risk and ambiguity[1]. Risk is uncertainty where the outcome of a decision is unknown, but the possible outcomes and their probabilities are known. For example, there are two possible outcomes of a fair coin flip: head or tails, and their probabilities are both known to be 50%. However, for decision-makers in the real world, the true underlying probability distributions are typically unavailable, or too expensive to calculate. Decisions in everyday life are much more reflective of ambiguity, than pure risk. Ambiguity is uncertainty where the outcome of a decision is unknown, and there is limited or no information about the possible outcomes or their likelihoods. For example, toy vending machines in grocery stores give us an idea of real-life ambiguity in decision outcomes. Perhaps the front of the vending machine shows examples of all of the toys inside, or perhaps it's showing just a few of the possibilities. Further, we have no idea how many of each of the possible toys there are in the vending machine – or in other words, we don't know the true probability of receiving one of these possible toys. It is of important note that risk, as formalized here, has nothing to do with the hazards of a choice. For example, some would say skydiving is 'risky' in the sense that there is an understood danger if anything were to go wrong. Here, risk is determined in relation to the probabilities of the possible outcomes. For example, a 50/50 gamble between two options is the riskiest, in the sense that it is the least certain you can be about what the outcome will actually end up being. On the other hand,

a 75/25 gamble between two option is less risky than a 50/50 gamble, because one of the outcomes has a higher likelihood. Further, ambiguity, as formalized here, has nothing to do with sensory discriminability; instead, the decision ambiguity discussed here relates to the information available about possible choice outcomes and their likelihood of occurring.

For hundreds of years, mathematicians, economists, psychologists, and other researchers across various fields have formalized the risk and ambiguity in everyday decision making, and how to best deal with this uncertainty. In addition, researchers have made note of the numerous observable paradoxical choices that individuals engage in, when trying to deal with risk and ambiguity. While a toy vending machine is a frivolous example of ambiguity in decision making, there are plenty of consequential decisions involving ambiguity in an individuals' life, such as choosing between possible universities or jobs, and in history. For example, in a press briefing in 2002, the United States Secretary of Defense Donald Rumsfeld utilized ambiguity as a fearmongering tactic to garner public support of the United States to invade Iraq and the Middle East. When asked about Iraq's possession of weapons of mass destruction and whether there were connections between Iraq and terrorist organizations, he stated, '…there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones'[2]. The point of this statement was to create fear around possibilities that we couldn't possibly know, and unfortunately, humans do not always act in logical ways in the face of ambiguity. Our current understanding of how humans and animals make decisions under risk and ambiguity, rationally or irrationally, can be described as economic decision theory.

1.1 **Economic Decision Theory**

Almost three centuries ago in 1738, Bernoulli described a coin flipping game where a player is offered a chance to win \$2 if a coin is flipped and turns up heads. Then, in the next stage, the player will win double the original reward, \$4, if the coin turns up heads again which has a 1/4th probability. In the next stage, the reward doubles again for an \$8 for another head, which has a 1/8th probability, and so on and so forth until the coin flip results in tails, at which point they will lose their initial entry fee and any prior gains. The puzzle lies in the question of, "What is a fair price to enter the game?" In order to determine this, we must consider what the expected value (EV) of this gamble would be. The EV can be calculated by multiplying the reward value by its probability of occurring, and taking the sum over all possible steps in the game. The issue is, that while the probability of flipping heads that many times consecutively shrinks, the payout is increasing at the same rate, as such the EV of the game is infinite, as follows:

$$EV = \frac{1}{2} \cdot \$2 + \frac{1}{4} \cdot \$4 + \frac{1}{8} \cdot \$8 + \frac{1}{16} \cdot \$16 + \frac{1}{32} \cdot \$32 + \cdots$$

$$EV = 1 + 1 + 1 + 1 + 1 + \cdots$$

$$EV = \infty$$

(Eq. 1)

Because this game has an infinite EV players should be willing to pay any price to play, according to *Expected Value Theory*, which was the accepted standard of the time. However, as Bernoulli outlined, individuals will not pay millions of dollars to play this game – which is irrational, if an individual is trying to maximize the EV. This became known as the St. Petersburg paradox[3], and Bernoulli deduced that while the EV may be infinite, the expected utility (EU) is not, and eventually plateaus as the player wins more money. Thus, the players' subjective gains in value,

or marginal utility, are not as large as they were in the beginning of the game – capping the value of the game to much less than ∞. Importantly, he also outlined how individual risk preferences can vary person to person, based on their current financial circumstances. This finding would form the beginning of the basis for modern economic decision theory, specifically expected utility theory.

Almost two hundred years later in 1921, Frank Knight elaborated on the idea of uncertainty, separating the idea of risk, or the known possible outcomes and the knowable probabilities of those outcomes, and 'uncertainty', or decision situations with limited to no information about the possible outcomes and their probabilities[4]. This important distinction was made to express the differences in decision making processes in regards to the two, specifically the inability of researchers to apply the concept of maximizing expected utility, like one would in risky decision making, in 'uncertain' conditions. In the decades following, mathematician John von Neumann and economist Oskar Morgenstern published their book, *Theory of Games and Economic Behavior*, where they expand on Bernoulli's initial concept of expected utility (EU) with a handful of axioms to explain rational decision making under conditions of risk[5]. Leonard Savage would further elaborate on individual preferences during decision making in the presence of risk, with what he described as subjective expected utility[6]. Both works introduce concepts that are the cornerstone for expected utility theory, game theory, and Bayesian statistics.

It would not be until 1961 when Daniel Ellsberg publishes "Risk, Ambiguity, and the Savage Axioms", where there was a real sense of investigating how decision-makers deal with 'uncertainty' – as it was known – as opposed to risk[1]. It is in this work that Ellsberg defines ambiguity, which was previously described as Knightian uncertainty[4], or uncertainty where the outcome is unknown and the possible outcomes and their probabilities are also unknown. In this

work, Ellsberg introduces a new scenario that displays paradoxical behavior that violates the prevailing theory of decision making under uncertainty, expected utility theory (EUT). The Ellsberg paradox describes a scenario where there is an urn with 90 balls, 30 are red, and the remaining 60 are either black or yellow. An individual is asked to bet on one of two gambles, (A) $100 if the ball drawn from the urn is red or (B) $100 is the ball is black. Then the individual is asked to choose between two other gambles, (C) $100 if the ball drawn from the urn is not black or (D) $100 if the ball is not red. In most cases, it has been shown that individuals will choose gamble A over B, and gamble D over C. The paradox lies in the fact that the people will choose to bet for known information – in this case gamble A that the ball selected is red, and gamble D that the ball selected is not red. This outcome violates Savage's sure-thing principle[6], requiring that decision-makers conserve their belief that red is more probable in the urn, which is consistent with choosing gamble A but not gamble D. This choice preference to choose based on known information, the 30 red balls we know about in the urn, instead of unknown information, or the unknown number of black balls in the urn, is known as ambiguity aversion. Importantly, decision-makers will avoid ambiguity at the expense of utility. This is a seminal work for understanding how risk and ambiguity effect choices.

Risk preferences and individual-specific decision patterns that belie expected utility theory are realized in the 1979 work of Kahneman and Tversky, "Prospect Theory: An Analysis of Decision Under Risk"[7]. It is in this paper that a value function is detailed and shown to be specific to changes in an individual's current situation, and their own specific preferences that are context dependent. The value function described is concave for gaining rewards, leading individuals to be risk averse for gains, and convex for losing rewards, or risk seeking. This collection of research spanning hundreds of years has informed our modern understanding of the rules of decision

making under risky and ambiguous conditions, and the many observable ways with which we break those rules in real life choices. While decision making can't always be accurately described by expected utility theory or prospect theory, they are the standard that researchers can apply to try to understand what rational economic decision making looks like under uncertainty, and how individuals can evaluate their options in complicated and perpetually changing environments.

1.2 **Reinforcement Learning**

Success in uncertain decisions rely on an individual's aptitude for navigating complex surroundings, evaluating many available possibilities, and approximating potential choice outcomes as closely to reality as possible. Further, integrating the outcomes of past choices to learn and inform future choices has a huge influence on individuals' success in receiving the highest value rewards. This is what is at the crux of reinforcement learning, simply put, it is how individuals learn to make the best decisions. Modern reinforcement learning (RL) has a past in behavioral psychology, specifically trial-and-error learning, optimal control of dynamical systems, and temporal difference methods[8].

1.2.1 **Trial-and-Error Learning**

Edward Thorndike was the first to describe the heart of RL, which is trial-and-error learning. Specifically, in 1911 Thorndike described the "Law of Effect", which defined a decision-maker's tendency to learn to repeat actions that were followed by good outcomes, and to stop actions that were followed by bad outcomes[9]. In other words, the *effects* of actions informed an

individual's future decisions of which actions to repeat and which ones to stop. This form of learning is different from just association, as it takes into account a history of past outcomes.

The term 'reinforcement' stems from Ivan Pavlov's work from 1927 describing classical, or Pavlovian, conditioning[10]. Pavlov was studying the physiology of digestion in dogs by measuring their salivation, when he discovered that the dogs would begin salivating not at the presentation of the food, but instead, when the experimenter entered the room to feed them. Pavlov followed this observation and discovered that he could elicit a physiological response (unconditioned response; UR) to stimuli that normally would produce no response, or a neutral stimulus (NS), by pairing it with a stimulus that naturally elicits the UR, or an unconditioned stimulus (US). The dog naturally salivating to the presentation of food would be the UR and the US, respectively. Then by pairing a sound prior to the food, the dogs would learn to associate the sound (conditioned stimulus; CS) with the presentation of food. Finally, after learning these associations, the sound (CS) would elicit the animal to begin salivating even without the immediate presentation of food – a conditioned response (CR) to the CS. Although reinforcement learning as we know it is more than learning by association, Pavlovian conditioning provided a critical framework for cognitive and behavioral research, and allowed researchers to utilize conditioning to investigate increasingly complex decision making paradigms.

Another key contribution that pushed trial-and-error learning towards reinforcement learning was that of Marvin Minsky in 1954. In his PhD dissertation, Minsky discusses computational models of reinforcement learning and an analog machine he created, with multiple components named Stochastic Neural-Analog Reinforcement Calculators (SNARCs)[11]. In the same year, Farley and Clark designed neural networks that learned through trial-and-error[12]. In 1963, John Andreae created a machine that used trial-and-error learning and included a model of

the environment, and later added a component to perform hidden state inference, or 'internal monologue' which he named STeLLA[13,14]. Around the same time as Andreae was creating his trial-and-error learning machine, Donald Michie was creating a system that could learn to play Noughts and Crosses (also known as Tic-Tac-Toe) called a Matchbox Educable Noughts and Crosses Engine (MENACE)[15,16]. These are just a few examples of the steps taken towards creating algorithms that utilized trial-and-error learning techniques that would inform modern reinforcement learning.

In 1975 John Holland introduced what would become a critical tool in reinforcement learning research, the *n*-armed bandit[17]. This mechanism was used to portray selection between some number (*n*) of independent 'arms', like a those of a slot machine, where learning takes place through choosing a different arm, receiving some outcome, then deciding whether to *exploit* the option you have already experienced, or *explore* alternate arms of the bandit. Holland's initial version of this concept was to utilize trial-and-error learning machines where algorithms learn to optimize genetic fitness by selection of biological variables such as mutation, crossover, and was based on an evolutionary adaptive system[18]. However, the key principles remain and are used in reinforcement learning research today. Holland also emphasized the importance of the dual control problem, introduced by Fel'Dbaum in 1965[19], which states the need to select a policy for behaviors of specific variables while simultaneously identifying other options in uncertain conditions. Holland specified the issue as one between exploiting known options and exploring new ones to gain information. Finally, in 1986 Holland introduced the idea of classifier systems, which were algorithms that use state-dependent rules to store information in order to create value functions associated with possible behaviors or policies[18]. This work would greatly influence Richard

Sutton, Andrew Barto, and Charles Anderson's work on the Actor-Critic reinforcement learning architecture[20].

1.2.2 **The Optimal Control of Dynamical Systems**

A second thread of the history behind reinforcement learning is optimal control theory; which deals with choosing a controller, or defined behavior of specific variables over states, to best operate a dynamical system, such that some objective output is optimized[21]. Richard Bellman developed an approach that involves solving a value function, or an optimal return function, to best optimize the dynamical systems output. Finding the solutions to these dynamical systems is also known as dynamic programming. In the case of reinforcement learning, a decision-maker (or agent) has the goal of optimizing the value accrued from their decisions. In order to achieve this objective, they must learn to behave optimally over states (i.e. learn an optimal controller) via successive approximations, or in other words, over trial-and-error learning.

Bellman also created a discrete, stochastic version of the optimal control problem known as Markov decision processes (MDPs)[22]. MDPs describe a situation where a decision-maker in a specific state can select an action that is available in that particular state, and then the decision-maker is moved to a new state and given a reward, much like Markov chains which describes a sequence of possible events where the probability of each event is dependent on previous states. The assertion of the MDPs is that these outcomes are partly random and partly under the control of the decision-maker. A policy is the specific action that a decision-maker is likely to make given a specific state. In 1960, Ron Howard introduced the policy iteration method for MDPs, where time steps are repeated over and over until the policy converges, giving an ideal action to take given the state, which is still used in modern reinforcement learning algorithms[23].

1.2.3 **Temporal Difference Learning**

Classical conditioning methods using a conditioned stimulus (CS), or secondary reinforcer, are critical in animal behavioral and cognitive research. Temporal difference (TD) facilitates learning by predicting future outcomes following a choice, and then evaluating whether said outcome is worse or better than predicted, and utilizing that evaluation update future predictions[11]. The point of this computation is to assist in learning of a secondary reinforcer, or CS, and correctly calculating what reward the secondary reinforcer predicts. Arthur Samuel was the first to create a checkers-playing artificial learning system that utilized TD-learning methods[24].

It was in the 1970's that Harry Klopf combined the trial-and-error learning theories with temporal difference (TD) learning, describing large adaptive systems that had multiple subcomponents that could locally reinforce each other while influencing the larger system as a whole[25]. Specifically, this system had excitatory input for rewards, and inhibitory input for punishments, in order to represent the 'hedonic aspects' of behavior. This assertion was critical for distinguishing reinforcement learning from supervised learning, which learns to implement choices based on generalizing from training examples instead of by trial-and-error. In their book, *Reinforcement Learning: An Introduction*, Sutton and Barto describe how Klopf's work would be a key influence in their integration of optimal control theory, trial-and-error learning, and TD-methods to build reinforcement learning, specifically using temporally successive predictions and the research into animal learning and behavior[8].

1.2.4 **Actor-Critic Architecture**

In 1983 Barto, Sutton, and Anderson would publish a method applying temporal difference (TD) methods to trial-and-error learning in what they called the actor-critic architecture (Fig. 1)[20]. In this model, the agent, or decision-maker, chooses a policy which creates an action that effects the environment of the agent. Then, after the action, the environment provides feedback in the form of reward and the state of the agent. The actor refers to the behavioral policy and the critic refers to the estimated value functions. The value function is the "critic" in that it critiques the policy selection of the actor, and updates it accordingly. Importantly, the TD prediction error is the sole critique from the value function, or the critic, and this is what is used to update the policy selection of the actor. This architecture is different from other methods by having a value function that is separate from the policy selection, or in other words, by having the critic and actor separated instead of being in the same function.

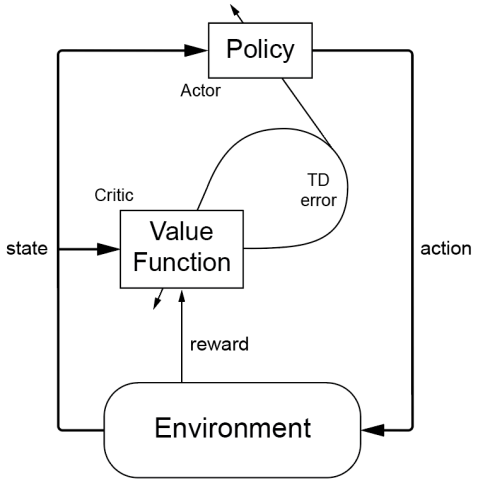


Figure 1: Actor-Critic Architecture of Reinforcement Learning
Reproduced with permission, Figure 6.15 of Sutton & Barto, 1998; Reinforcement Learning: An Introduction, published by MIT Press.

The original 1983 paper utilizing the actor-critic method used it to address Michie and Chamber's pole-balancing task, which was one of the earliest examples of a learning task under conditions with incomplete knowledge, or uncertainty[20,26]. The goal of the pole-balancing system is to move a cart left and right in an attempt to balance a long, hinged pole attached to the middle, and prevent it from falling over. Some important variables in this example are the angle of the pole and the location of the cart on its finite track. The state could describe situations where the cart

was towards the left, center, or right of the track, the velocity of the cart, and the direction and angle that the pole is leaning. The actor must select a policy for all of these states where the action selected exerts enough opposite force to straighten the pole back out, or keeps it steady and balanced. After a policy, and in turn an action, is selected by the actor, the environment will update the state, and return a reward, for example +1 for keeping the pole from falling for another step, or nothing for letting it fall. The value function is updated by the reward and the TD prediction error is sent by the critic, which updates the policy selection of the actor for future states. This model of reinforcement learning is still vital to neuroscience, machine learning, and psychological research done today.

1.2.5 ***Q*-Learning Algorithm**

The integration of trial-and-error learning, temporal difference learning, and optimal control, was when *Q*-learning, a model-free reinforcement learning (RL) algorithm, was introduced by Chris Watkins in 1989[27]. *Q*-learning is model-free, in that it does not require a model of the environment of the agent, and it learns the optimal policy that maximizes value in a finite Markov decision process. Importantly, *Q*-learning is an off-policy method, unlike Actor-Critic methods, meaning that the system learns the optimal policy independent of the agent's action selection process. In other words, the Actor-Critic methods of RL are on-policy, because it is only estimating the value function of the policy currently being followed. The *Q*-learning policy relies on value function, "*Q*", that the algorithm computes which returns the expected rewards for an action taken in a particular state. A convergence proof of *Q*-learning was given by Watkins and Peter Dayan in 1992[28]. Integrating *Q*-learning into the current methods is a critical step towards integrating optimal control, and temporal difference methods into RL.

1.2.6 **Modern Reinforcement Learning**

To make optimal decisions, an individual must learn to act in ways that provide them with the most rewarding outcome – which is subjective to that individual. While economic decision theory instructs how individuals choose based on their subjective preferences, reinforcement learning tells us how an individual learns what exactly the best outcome is to them. Reinforcement learning is incredibly important in a complex and uncertain world, where things are more nuanced than just worse or better, and learning about uncertain environments helps us get more value in the long run. Imagine that I want to go out and buy a coffee. Redhawk Coffee is a favorite coffee shop of mine, and I have accrued hundreds of experiences buying a coffee there, and then experiencing some outcome in terms of the subjective value that coffee provides for me. Every time I receive a coffee from Redhawk, I taste it, and experience a prediction error based on the subjective value I experience from that particular cup of coffee. This is a temporal difference (TD) learning method used to update the value of options in the everyday decisions we make. Sometimes my cup of coffee is worse than my average experience, and sometimes it's better. I learn from all of these experiences, and they accumulate to create a distribution of outcomes that build my average expectation of the value of a cup of coffee from Redhawk. In this instance, the outcomes are normally distributed around my average expected value (EV) of a coffee at Redhawk (Fig. 2, left panel). While this alone is a useful mechanism to build an accurate expectation for the value of receiving a coffee at Redhawk, it is most useful in helping us determine what decision to make in the first place. Most often, we make decisions between multiple uncertain choices, and this is

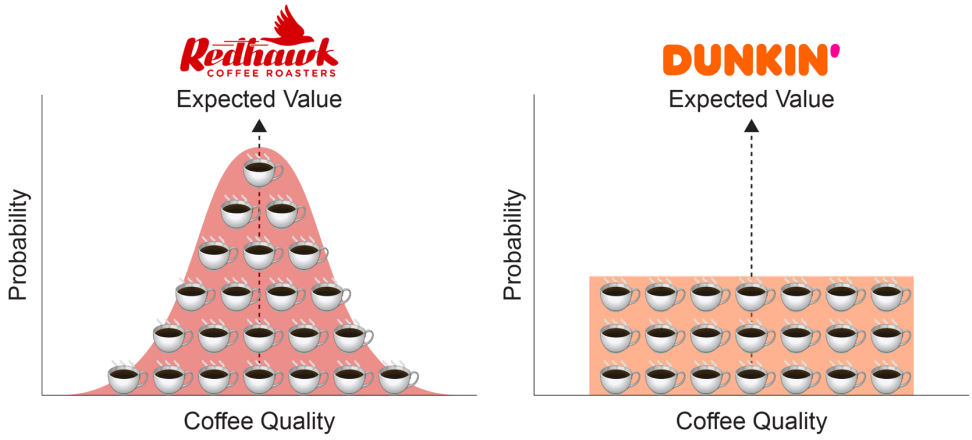


Figure 2: Experiences Shape Value Expectations
Probability density functions representing the quality of a cup of coffee for Redhawk Coffee Roasters (left), and Dunkin' Donuts (right).

where our cumulative experiences help us make decisions. While I frequent Redhawk, I sometimes go to Dunkin' Donuts for coffee. However, unlike my distribution of experiences at Redhawk, my experiences at Dunkin' Donuts are much less consistent and are uniformly distributed around an average subjective value (Fig. 2, right panel). This means that even though my average subjective value of coffee from both of these places is similar and is across the same range of worse or better than average outcomes, a cup of coffee from Redhawk is much more consistent, and I am more likely to experience a much worse than expected – or much better than expected – cup of coffee from Dunkin'. What does this mean for me as a decision-maker? Well, if I prefer the reliability of the quality of the cup of coffee at Redhawk, then I might choose to go there as to not risk the possibility of a getting a much worse than expected cup of coffee at Dunkin'. On the other hand, perhaps I am feeling optimistic about my chances of getting a better than expected cup of coffee at Dunkin', and I take the risk and choose to go there instead of Redhawk, as I am much more likely to a much better than average cup of coffee at Dunkin' based on my distribution of past experiences. Every one of these outcomes provided a TD prediction error, and allow me to update what to expect when I choose to buy coffee from one of these places. This is a simple example of how distributions over

rewards influence reward-based decision making in uncertain environments that is happening in our everyday lives.

Reinforcement learning is, put simply, the updating of values of behaviors, and these values should be considered subjective to the individual and can be influenced by the uncertainty surrounding the options and their outcomes. Combining these concepts, reinforcement learning and economic decision making, are the heart of this body of work and specifically, the effects of uncertainty on the brain and the neural systems involved in learning and decision making. The next section will be a discussion of function of multiple regions in the brain that mediate decision making through their representations of reward prediction errors, value, and uncertainty.

1.3 **How the Brain Mediates Reward Based Decision Making and Learning**

Reinforcement learning (RL) has become a universally accepted way to consider how humans and animals learn from experience and make appropriate decisions. Much of early RL research utilized these computational concepts to create artificial learning systems. At the same time, because of the computational context it provides, neuroscientists have sought out how neurobiological systems fit into this framework. A key factor in RL is value functions, and economic decision theory has given clear standards on how individuals construct value (utility) functions based on their subjective experience in an uncertain world. The following sections will outline different key brain regions that perform uncertainty and value coding, and mediate learning. A caveat to keep in mind is that a lot of these brain regions perform multiple different roles depending on the situation[29]. These brain regions do not exist to perform in a singular context,

and as such, most need to, and will, dynamically adapt to fit the role necessary to successfully learn and optimize performance in different behavioral conditions.

1.3.1 **Neural Coding of Reward Prediction Errors**

Reward prediction errors (RPEs), in terms of temporal difference (TD) learning rules, are the difference between the received minus predicted reward:

$$RPE = Reward\ Received - Reward\ Predicted$$

(Eq. 2)

The predicted reward is a point estimate that represents the expected subjective value. In regards to reinforcement learning (RL), this acts as the critic's signal back to the actor. This signal can be used to update the state value estimate, and be integrated into the policy of what decision to make in future iterations. The first neural substrate to become widely associated with coding RPEs are midbrain dopamine neurons. Before dopamine neurons were known to code RPEs, many scientists studied their role in generating movements, due to the motor impairments resulting from lesion studies and evident in Parkinson's disease, which is a result of the deterioration of dopaminergic neurons[30-34]. However, following multiple investigations to uncover the role of dopamine neurons, it was in 1997 that Schultz, Dayan, and Montague discovered that dopamine neurons code reward prediction errors[35]. They saw that dopamine neurons showed phasic activations to receiving unpredicted rewards (Fig. 3, top). Further, when they associated a cue with a reward, dopamine neurons stopped having phasic activations at the time of reward, but instead became activated at the cue presentation that fully predicted that a reward would be given (Fig. 3, middle). In addition, if the reward was withheld following the reward-predicting cue, dopamine neurons showed phasic

pauses in firing (Fig. 3, bottom). This finding perfectly capitulated a TD prediction error (Eq. 2). It is important to note that RPEs happen not only following actual rewards – or primary reinforcers – but also following secondary reinforcers, like a conditioned stimulus (CS) that predicts reward. To contextualize how real world experiences can produce the firing patterns shown in Figure 3, imaging you are at a restaurant and you have ordered your food and you are waiting for it to come out. In the top panel of Figure 3, there is no CS, so no RPE (dark blue), followed by an unpredicted reward, causing a positive RPE (light blue) – or a phasic burst in firing. This might happen if you don't see the waiter coming towards you with your food and they just set it down all the sudden. You didn't see the waiter coming and thus had no prediction of food being delivered, so when you unexpectedly receive the food you experience a positive RPE. In the middle panel, there is a CS that predicts a reward, causing a positive RPE (dark blue), followed by the predicted reward, causing no RPE (light blue). In this instance, you see the waiter coming out with food, and you have the positive prediction error because you went from no food to food is on the way, and delivery was fully expected, resulting in no prediction error. In

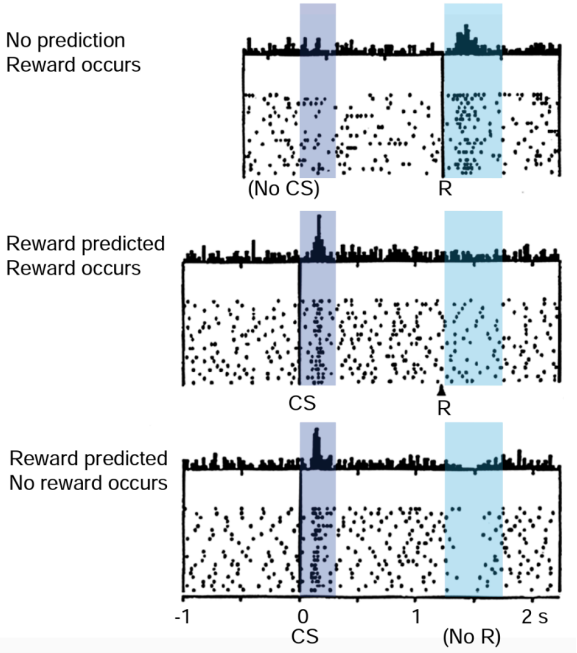


Figure 3: Midbrain Dopamine Neurons Code Reward Prediction Errors

Raster plots and peri-stimulus time histograms (PSTHs) from extracellular recordings in midbrain dopamine neurons. Top panel is aligned to reward (R), middle and bottom panels aligned to conditioned stimulus (CS) presentation. **Top**, No CS and thus, no reward prediction occurs (0 RPE, dark blue), followed by an unpredicted reward (Positive RPE, light blue). **Middle**, CS predicts reward (Positive RPE, dark blue), followed by the predicted reward (0 RPE, light blue). **Bottom**, CS predicts reward (Positive RPE, dark blue), followed by the predicted reward being withheld (Negative RPE, light blue). Dark blue shaded regions indicate RPEs to secondary reinforcers (the CS or no CS), light blue shaded regions indicate RPEs to primary reinforcers (juice or no juice reward). *Reproduced and modified with permission, based on Figure 1 of Schultz, Dayan, & Montague, 1997; published in Science by the American Association of the Advancement of Science.*

the bottom panel, there is a CS that predicts a reward, causing a positive RPE (dark blue). However, this time, no reward is given, causing a negative RPE (light blue). In this instance, you see the waiter coming out with food so you have a positive RPE, but the waiter walks past you with the food to a different table, so instead of getting food like you expected, you get nothing, causing a negative RPE.

Waiting for food at a restaurant is an intuitive example of real-life prediction errors. However, a lot more than reward vs. no reward goes into how our neurons calculate the "Reward Predicted" aspect of the RPE equation (Eq. 2). Consider a monkey is in an environment where the average reward they can expect is 0.6 ml, based on a cue that predicts equal probabilities of receiving 0.4 or 0.8 ml of juice. When they see the cue, they have a RPE that reflects the point estimate of the subjective value of that cue, which is, 0.4 ml*(*P*)0.5 + 0.8 ml*(*P*)0.5, which results in an expected value of 0.6 ml for the cue that predicts those outcomes. Now, when they receive either the 0.4 or the 0.8 ml reward, there will also be a RPE response. A negative RPE response if they receive the 0.4 ml of juice, as it is worse than predicted (0.4 ml < 0.6 ml) – or a positive RPE response if they receive the 0.8 ml of juice, as it is better than predicted (0.8 ml > 0.6 ml). The absolute reward magnitude is not the only consideration when calculating the expected value of a cue. The value received from a reward is not the same across individuals, and as such, the estimates of subjective value that are used by dopamine neurons to calculate reward prediction errors also reflect individual differences. Utility is defined as the unique worth or benefit that a person gains from a reward, and dopamine neurons use utility to code reward prediction errors[36]. The utility of an outcome incorporates any kind of feature that could alter how much a person values it, and dopamine neurons have been shown to incorporate them individually as well. Different features of reward outcomes, such as reward type, volume, delay, effort required, and probability are all

integrated on a common scale of utility[37-41]. In addition to incorporating all of these facets of reward, dopamine neurons also adapt in order to most effectively code reward prediction errors regardless of the absolute utility, such that RPEs in a context with rewards of small utility are efficiently coded for maximal learning, and RPEs in a context with rewards of large utility are also efficiently coded[42]. This is important because even though the utility of a coffee is much smaller than that of a car, it is necessary to adapt prediction error coding in order to learn which outcomes were the best in both contexts. To elaborate on this, if you want to buy a car and you go to test drive a few, you will experience RPEs in relation to whether the experience driving them is better or worse than you were expecting, and these RPEs will utilize the maximal range of firing rates in dopamine neurons. Then, if you go and buy a coffee, the comparative utility is much smaller, but you still want to effectively evaluate if it was better or worse than you were predicting. If dopamine did not adapt to different contexts, the RPEs in response to the much smaller utility of a coffee would be so minimal that it would become a useless mechanism. However, because dopamine neurons can adapt to different rewarding contexts, they can still utilize their full firing range, even with smaller utility decisions, like a coffee. Reward prediction error signals from dopamine neurons have been proven to be causal in effecting behavior. Specifically, animals will choose a stimulus that predicts optogenetic stimulation of dopamine neurons over a choice with the same juice reward but no optogenetic stimulation[43,44]. Furthermore, if dopamine neurons are inhibited, this outcome is reversed, and behaviors that predict dopamine inhibition are avoided[45,46]. Thus, RPE signals can causally influence how we learn and make reward-based decisions.

Two other nuclei that perform reward prediction error coding are the lateral habenula (LHb) and the rostromedial tegmental nucleus (RMTg). These regions both provide input to dopamine neurons – specifically, a small amount of direct afferents to dopamine from the LHb[47,48],

along with additional input routed disynaptically through the RMTg[49,50]. The LHb plays a key role in learning from non-rewarding or even aversive events, by coding an inverse reward prediction error where neurons are activated by non-rewarding or aversive stimuli, and silenced by reward and reward-predicting stimuli[51,52]. The RMTg also codes inverse reward prediction errors[49,50]. While the RMTg receives inverse reward prediction error signals from the LHb, the anterior cingulate cortex also inputs to RMTg and provides a teaching signal to learn to avoid aversive outcomes[53]. Stimulating the LHb facilitates behavioral avoidance[54,55]. Further, there is a direct causal relationship between the LHb and dopamine neurons, such that activating the LHb causes inhibition of dopamine neurons, which is mostly mediated through the disynaptic connections through RMTg[56-59]. Further, ablation of the LHb prevents learning from unrewarded stimuli, and even minimizes animal's ability to create nuanced behavior and graded dopamine RPE responses to stimuli with varying expected values[60]. The LHb, RMTg, and midbrain dopamine neurons all code a form of reward prediction error, and provide a neural teaching signal to instruct which behaviors to repeat and which to stop. These teaching signals are utilized in downstream brain regions to update value representations of stimuli and actions that preceded the reward prediction errors.

1.3.2 **Value Representations in the Brain**

A critical part of economic theories of decision making is that decision-makers often value the same options differently given their own conditions, and when compared to other individuals. The circumstances the person is under, their subjective preferences, their internal states, and risk tolerance all play into how an individual evaluates possible choice options. Devaluation is a classic example showing that internal state can change subjective preferences through selective

satiation[61]. Consider a person who prefers orange juice over apple juice. After a week of having orange juice every day, they choose apple juice over orange juice. Wealth can also modulate individuals' risk attitudes and consequently, their decisions. Consider two individuals go to a casino. They both play blackjack and both have won $10,000. One of these individuals is already rich, so they continue to play and gamble, as they are not satisfied by the $10k of winnings, and they are not worried about the possibility of losing all their winnings by continuing to gamble. The other individual, however, is poor and sees $10k as a huge win, and chooses to cash out in order to avoid the risk of losing what they have already earned. These examples suggest that economic decision making is a dynamic process and our brains must represent and integrate multiple factors across different domains during economic decision making.

1.3.2.1 **Subjective Value Coding in Cortical Regions**

One of the most well-characterized regions for subjective value coding is the orbitofrontal cortex (OFC), which is located at the most ventral portion of the prefrontal cortex. The past two decades of research in the OFC has shown how essential this region is for encoding subjective value of choice options, regardless of the action taken by the subject. These findings indicate a role for OFC in instructing downstream regions which options are of the highest value to the decision-maker. Interest in OFC function stemmed from the observable effects of frontal lobe damage in humans, and subsequent lesion studies in monkeys. For example, macaque monkeys with OFC lesions show a variety of behavioral impairments in multiple different tasks. One striking example reveals the loss of ability to update the value of a previously non-rewarding object[62]. In the first study characterizing the electrophysiological functions of OFC neurons in awake and behaving macaques, authors noted that neurons did not code information relevant to the performance of the visual discrimination task, but indeed, were very active in response to the

different stimuli and their values[63]. Further, they made note that some neurons of the OFC would show the same level of activation for a particular reward at the time of the visual stimulus that predicted it, and at the reception of the reward itself; a finding that would foreshadow our current understanding of subjective value coding in the OFC. In another study, OFC neurons stopped responding to rewards that had been devalued, showing that as the individuals' subjective value of a reward decreased so did the coding of that reward in OFC neurons[64]. In 1999, Tremblay and Schultz recorded from OFC neurons in macaque monkeys performing a delayed-response task, where the monkeys had to move their hand to a lever on the side that a previously shown image had appeared[65]. The task was set up in blocks where two different rewards could be given, depending on the identity of the visual stimuli. In blocks of trials where the monkeys' preferred (Reward A) and second preferred reward (Reward B) could be received, OFC neurons showed higher firing rates following Reward A. However, when they were in blocks of trials where their second (Reward B) and third preferred reward (Reward C) could be received, OFC neurons showed the same pattern of firing as in blocks with Rewards A and B, but in this case, OFC neurons showed higher firing rates following Reward B. This result indicated that the OFC coded the relative value of the rewards available during a specific block and illustrates the flexibility of OFC neurons – they can update their coding scheme based on the options that are available in a given environment, much like range adaptation seen in dopamine neurons[42]. More recent studies have shown that OFC coding meets two important conditions required to be considered an abstract representation of subjective value – or in other words, in the space of economic 'goods'[66]. First, the coding of value must be independent of the sensorimotor contingencies necessary for making a choice; and second, the coding should encompass all external (i.e. reward identity and amount, delay to reward, effort required, uncertainty) and internal (i.e. attitudes towards uncertainty,

motivation, delay tolerance, etc.) aspects that could be integrated to compute the value of the option. These conditions have certainly been met, as many studies have substantiated that the OFC codes subjective value of choice options, integrating multiple relevant aspects of the option; and that this coding was independent of the subsequent choice and behavioral action the animal planned and consequently completed[66-77]. Further, decoded activity of neural populations in the OFC shows the value comparisons between two choice options. Specifically, when the decision is fast, the coding between two options separates quickly; when the decision is slow and the individual takes more time to deliberate, the separation between the coding of the two choice options is smaller[78]. This is a reasonable result, as an individual may not be as decisive when they value two options similarly, and it should not take an extensive amount of consideration to choose an option that is subjectively much more valuable. It has been suggested that the location of the OFC in the prefrontal cortex is quite distinct, as it receives inputs from visual, olfactory, gustatory, and visceral sensory areas[70,79]. This junction of information in the OFC provides the neurons in this area a unique ability to integrate multiple sensory modalities to create a comprehensive subjective valuation of choice options in the environment.

In addition to the OFC, there are a number of other prefrontal cortical regions that perform value coding. The medial prefrontal cortex (mPFC) has been shown to encode the value of a chosen option, such that the activity persists following selection of the option – which could act as a remedy for the credit assignment problem[80] by allowing memory of the value of the selected action to remain available[69]. Further evidence to support the mPFC role in credit assignment comes from a choice task where two options are individually and serially presented to monkeys. In this task, mPFC neurons preferentially encoded the value of the first option presented and persisted during and after receiving the reward[72]. In a choice task where humans made choices between immediate

and delayed monetary rewards, functional magnetic resonance imaging (fMRI) blood-oxygen-level dependent (BOLD) activity in mPFC was correlated with the subjective value of the rewards in a common currency that integrated the delay and reward volume for multiple options[81]. Other fMRI studies have shown that ventromedial prefrontal cortex (vmPFC) activity also integrated risk and ambiguity[82], and gains and losses[83] into their subjective value estimates. The lateral PFC (lPFC) has been shown to encode stimulus identities as well as the location of option that will be chosen, indicating a role for this region in transforming goals in the environment into actions that will allow us to obtain these goals[84-86]. Single-unit electrophysiology in the dorsolateral PFC (dlPFC) – specifically Broadmann area (BA) 46 – of rhesus monkeys performing an oculomotor delayed-response task showed enhanced activation in the delay before making a saccade to a larger reward-predicting target stimulus in their response fields[87]. Other work has shown action-value functions in this region as well, informing which actions had preceded the best rewards[88,89]. In more recent research, decoding of neural activity in the dlPFC has shown value representations for novel pairs of rewarding choice options, and further, that these representations begin to appear as the values of the stimuli are learned and updated following a value reversal[90]. In addition, this study also demonstrated that the dlPFC showed expanding and unique coding space for the numerous individual pairs of reward-predicting stimuli, signifying coding of specific object identities and not just their value. Indeed, the dlPFC has been shown to be flexible in the variables it encodes dependent upon the demands of the task – a concept known as 'mixed selectivity' due to the adaptable rules of coding seen in dlPFC[91].

Other regions of the cortex, such as the cingulate cortex, and posterior parietal cortex have also been implicated in coding value. The anterior cingulate cortex (ACC) lies on the ventromedial wall of the frontal lobe and has been shown to code a state value function, like that seen in a

reinforcement learning algorithm[92]. The posterior portion of the cingulate cortex (PCC) also shows value related signaling, specifically, neural signals about value differences between the actual reward received and the average reward rate in an environment are carried into the next trial, essentially keeping track of a previous trials' risk to potentially inform future decisions[88,93,94]. In the parietal cortex, similar value functions related to the action or state value have also been reported. Many neurons in the lateral intraparietal cortex (LIP) encode value functions for eye movements[95-99]. However, more recent work has shown the LIP encodes the salience of choice options as it relates to the magnitude of value gains and punishments, as opposed to just value[100]. This suggests LIP can mediate a wider scale of outcomes than previously suspected, which is critical when making decisions where there may be a tradeoff of positive and negative outcomes. These are just a few key regions of the cortex that mediate value into their neural coding.

1.3.2.2 **Subjective Value Coding in Subcortical Regions**

One subcortical portion of the brain, the basal ganglia, are comprised of the striatum, pallidum, and the substantia nigra[101]. The basal ganglia are classically thought to be a key controller of movement, due to the degradation of motor control seen in Parkinson's disease following the degeneration of midbrain dopamine neurons and consequent dopaminergic denervation of the sensorimotor striatum[30,102]. Observations in disease states lead to the characterization of the "direct" and "indirect pathway" through the D1- and D2-medium spiny neurons (MSNs), which appear to mediate opposing effects on movement[103,104]. However, research has shown that while basal ganglia projections mediate the control of movement, there are also substantial contributions to integrating value and learning signals from multiple regions across the brain. Characterizations of the anatomy and functional roles of the basal ganglia show distinct, parallel basal ganglia-thalamocortical circuits[105]. More recent research has shown the role of the

reciprocal connections between the cerebellum and the basal ganglia as well – suggesting multiple nodes in an integrated network to mediate behavior[106,107]. These anatomically distinct yet functionally related subcortical networks integrate information from motor, limbic, and cognitive networks to produce sophisticated behavior; and reward-mediated decision making is included in these processes.

The striatum is often credited as the critical mediator in reinforcement learning (RL) through learning and encoding of value[108]. The ventral striatum (VS) consists of the nucleus accumbens (NAc) and the olfactory tubercle (OT), while the dorsal striatum (DS) consists of the caudate (Cd) and putamen (Pt). Both of these regions receive vast input from midbrain dopamine neurons[109], which relay reward prediction error responses to instruct which behaviors to repeat and which to stop[110,111]. Much of RL-based research in the striatum separates into action-value and stimulus-value, and some have postulated dissociable roles of the DS and the VS[112]. Multiple studies recording across the entirety of the striatum, in both phasically active neurons (PANs, putatively MSNs), and tonically active neurons (TANs; putatively cholinergic interneurons) have shown reward-related signals in decision making tasks[113-118]. In the DS, a seminal study showed that the value of two separate actions were learned and encoded in MSNs in the Cd and Pt of macaque monkeys[119]. Specifically, the values encoded represented the learned action-values that incorporated the history of outcomes that reflected the probability of choosing a particular action, as estimated by a RL model. However, the DS is not limited to encoding only action values. Microstimulation in the Cd of the DS causally increases the value of specific choices, and the probability of choosing a specific option, regardless of the actions necessary to make the choice[120]. This finding refutes the idea that the DS is limited to perform as the RL 'actor', and speaks to the larger role of the DS as a mediator of learning both action and stimulus values.

In terms of the ventral striatum (VS) and its role in reinforcement learning (RL), some have described the role the VS played to be one of the critic, incorporating error signals from midbrain dopamine neurons to the value function[112]. However, lesions to the VS in monkeys produces specific deficits in assigning and updating value to rewarding stimuli, but not to rewarding actions[121]. If the VS were the critic in a RL framework, there would be deficits to both stimulus- and action-value learning, as it should learn value regardless of the modality of the association. The VS receives dense projections from midbrain dopamine neurons, and as such, reward prediction error (RPE) signals can be detected in this region. Specifically, blood oxygenation level dependent (BOLD) responses from MRI studies show RPE signals in the VS that mimic those seen in the midbrain[122]. In addition, dopamine concentration in the NAc also reflects RPE coding seen in midbrain dopamine neurons, suggesting a highly conserved electrochemical RPE signal from soma to dopamine release at the terminals in the NAc[123]. The neurons of the VS, however, integrate a multitude of different facets regarding rewards into their neural reward signals. Putative MSNs of the VS encode the value of rewarding options, and even hold a representation of the outcome of previous choices, perhaps to optimize value estimates and decisions on subsequent trials[124-128]. The role of the VS is minimized in deterministic choices, and is only really essential in probabilistic RL environments to stabilize value representations[127,129,130]. Perhaps this is due to the minimal learning necessary for deterministic environments, and thus, the lack of necessity for a region that receives such dense innervation from dopamine neurons, and the RPE learning signals they send.

Another region of the basal ganglia that is a critical mediator of reward based decision making is the pallidum[131,132]. The internal segment of the globus pallidus (GPi) plays a critical role in basal ganglia-thalamocortical circuits. A major output center of the basal ganglia, the GPi sends

motor and non-motor signals to thalamic and other brain stem regions[101,133,134]. In addition to playing a critical role in movement, the GPi also projects to the lateral habenula (LHb), which computes an inverse reward prediction error. The GPi connections to the LHb send information about expected reward following target appearance[135,136]. The external segment of the globus pallidus (GPe), also encodes reward related information, specifically, an integration of the future movement to perform and the expected reward following that movement[137]. Finally, the ventral pallidum (VP) has been postulated as a limbic output region of the BG, due to its connections with canonically limbic regions such as the VS, amygdala, and OFC[138]. While typically thought to translate reward-based motivation to movements necessary to acquire rewards[139], the VP also encodes preference for different rewarding stimuli[138]. These BG regions and their role in reward coding and decision making reflect the previously mentioned multiple parallel cortico-basal ganglia circuits that mediate multiple necessary functions in the brain[101].

While the striatum is typically thought to be the main subcortical region that learns and encodes values that drive decision processes, the amygdala has also been shown to perform similar value coding. The amygdala is a heterogeneous nucleus located in the temporal lobe just beneath the uncus that also receives dense dopaminergic projections from the midbrain[140,141]. The amygdala is historically known for its role in emotion and Pavlovian learning[140,142-145]. However, decision making research has shown that it also has a prominent role in reinforcement learning. Studies have shown that the amygdala encodes a state value function[146,147]. Further, distinct subpopulations of the amygdala encode either positive or negative value[148]. Interactions between the PFC and the amygdala are critical for reward-guided behavior, and disrupting this connection through amygdala lesions significantly impairs OFC's ability to encode reward value[149,150]. Similar to the neural signals seen in the VS, the amygdala encodes information about a preferred stimulus,

including its value and identity, and this signal is held across trials[84]. The amygdala has also been shown to encode information about the immediate value of exploiting known rewards in addition to the possible future value of exploring unknown rewards, suggesting a key role in explore-exploit decision making[127]. There are both dopamine D1 and D2 receptors in the amygdala, and dopamine has been shown to be necessary to make value representations labile in order to update them following dopamine reward prediction errors[151,152]. Further, increased dopamine availability through pharmacological interventions increases the likelihood of exploring novel choice options[153]. While the amygdala and the VS seem to have overlapping roles in reward-based decision making, consider the security of having multiple regions perform similar computations in the case that one of these regions were to fail in its role from either damage or disease states. Further, studies in monkeys have shown that animals with VS lesions are less impaired in RL tasks than animals with amygdala lesions, insinuating that while these regions have similar roles in value coding, there are still differences in their overall roles in RL, perhaps specifically through the effect they have on regions they project to[129].

Similar to the basal ganglia, the cerebellum is commonly associated with movement, specifically motor learning and movement coordination[154]. Specifically, about the role of the cerebellum in error-based learning in regards to motor movements. In the formalized Marr-Albus theory of cerebellar learning[155], parallel fibers from granule cells have a representation of the state of the system as well as the current motor commands (known as an efference copy). Comparatively, climbing fibers stem from the inferior olive nucleus and compute a motor-based prediction error, and both climbing and parallel fibers synapse on Purkinje cells. However, more work has shown the role of the cerebellum in reward-based decision making and learning. One study using two-photon calcium imaging showed granule cells responded to reward anticipation,

while others responded to reward or reward omission[156]. Further, neuroimaging studies have shown reward and reward prediction error signals in the anterior portion of the cerebellar cortex[157]. To summarize, similar to the parallel thalamo-cortico-basal ganglia loops that show topographic functional specificity, the cerebellum is also interconnected to the BG, with specific topography related to motor, cognitive, and limbic functions[158,159]. These reward-related signals across vast cortical and subcortical regions of the brain are critical in reward-based decision making.

1.3.3 **Uncertainty Coding in the Brain**

Uncertainty around reward outcomes, and the subjective value of the cues that predict those outcomes, are incredibly difficult to disentangle due to the fact that subjective value integrates reward uncertainty to create value estimates. In order to code subjective value – like in the previously detailed brain regions – there must be some biological representation that can be used in these calculations. While there are many regions that have been shown to reflect value, there are much fewer instances of uncertainty coding across the brain. A few regions that have shown uncertainty-dependent neural coding include the anterodorsal septal nucleus, striatum, multiple prefrontal cortex regions, parietal, and cingulate cortices.

From fMRI studies in humans in choice tasks, there has been evidence of a few regions with uncertainty-specific activation. The inferior frontal gyrus activity following safe and low risk options correlated with higher risk aversion in subjects, while the striatum and cingulate showed activations in response to making more risky choices – and these combined signals were able to predict choices made in risky decisions[160]. One study showed the medial prefrontal cortex (mPFC) correlated with amount of ambiguity in the environment and ambiguity aversion[161]. Another showed subjective value activations in ventromedial prefrontal cortex (vmPFC), while the

dorsolateral prefrontal cortex (dlPFC) was specifically activated by the level of uncertainty in those subjective value predictions[162]. Further, dlPFC manipulations with transcranial magnetic stimulation was shown to modulate risk preferences[163,164]. Uncertainty in subjective value estimates could be interpreted as ambiguity – as the uncertainty could be attributed to missing or unclear information in the underlying statistics of a choice environment. Indeed, another study showed distinct regions for ambiguity and risk, where the lateral prefrontal cortex (lPFC) showed modulations specific to ambiguity preference, while the posterior parietal cortex showed risk preferences[165] – suggesting two dissociable circuits for the two forms of uncertainty.

In electrophysiology experiments, results have shown distinct populations of orbitofrontal cortex (OFC) encoded reward value, and uncertainty as a function of variance in the possible outcomes[166,167]. Another study showed neurons in the macaque posterior parietal cortex coded uncertainty in terms of the coefficient of variance of rewards[93] – similar to activations seen in fMRI experiments activations based on risk preferences[165]. As mentioned previously, midbrain dopamine neurons showed graded increases of baseline firing rate dependent upon the uncertainty of reward outcome in risky cues[40]. In addition, the primate dorsal striatum has been shown to code risk in stimulus-outcome associations during decision making, with firing rates being higher for cues with higher risk[168]. Finally, in a study recording from the anterodorsal septal nucleus in rhesus macaques, single neurons showed uncertainty coding that reflected the inverted U-shape that describes the amount of uncertainty in terms of Shannon entropy of risky decisions (Fig. 12) – which was completely independent of the expected value of the gamble. These results provide ample evidence that uncertainty is relevant and elicits different responses across the brain, and can represent uncertainty in general, and even individual preferences for both risk and ambiguity.

1.3.4 **Ubiquitous Reward Modulation to Optimize Behavior**

In the previous sections, we discussed critical brain regions that perform reward-based decision making and learning. However, many regions that are primarily associated with other behavioral functions are still modulated by value. This pervasive reward modulation across the brain is critical in attenuating behaviors that are more rewarding than others. Consider some of the 'primary' regions of the brain: the primary visual cortex (V1), auditory cortex (A1), and motor cortex (M1). The primary visual cortex (V1) is located in the most posterior portion of the occipital lobe, and is known to be organized into cortical columns, and the receptive fields of V1 neurons are tuned by ocular dominance and line orientation[169-171]. However, there has been research that shows that rewards can modulate coding in this region. Specifically, it has been shown in that reward modulates the firing rate in firing of V1 neurons of rats[172] and rhesus monkeys[173], and in V1 fMRI responses in humans[34]. In one study, V1 neurons in rats were shown to fire in a way that coded the timing of rewards with different delays following a visual stimulus[172]. In another study, neurons in the rhesus monkey V1 showed heightened responses following stimuli that predicted higher value rewards[173]. These results demonstrate that rewards influence the neural activity in V1, a region that is critical in early visual processing.

The primary auditory cortex (A1) is positioned in the superior temporal gyrus of the temporal lobe, and is characterized by having a tonotopic map of sounds, where low to high frequency sounds are coded anterior to posterior[174-176]. Research using classical conditioning paradigms with rewarded and unrewarded tones has shown that the best frequency of the neurons in A1 shifts towards the rewarded frequency, but not the unrewarded one[177,178]. In addition, in an electrophysiological mapping study of neurons in A1 of owl monkeys, rewarded tones had enhanced spatial representations of rewarded tones in the tonotopic map, leading to enhanced tone-

discrimination performance[179]. Further, in a task where high and low reward was signaled by a visual cue, then rhesus monkeys had to withhold response during the auditory Wait signal, and then respond at the auditory Go signal, fMRI showed enhanced activity in A1 during the Wait signal, even though the tone was identical[180]. Interestingly, in a study in gerbils, the authors found that optogenetic stimulation of ventral tegmental area (VTA) dopamine neurons enhanced frequency-specific gain amplification in A1, and this enhancement was specific to the A1 cortical layers that receive direct innervation from dopamine neurons[181]. VTA dopamine neurons are known for reward prediction errors and sending a neural teaching signal to indicate which rewarding behaviors to repeat and which behaviors to stop. This research shows that this dopamine teaching signal regarding rewards extends to primary sensory areas like A1, and teaches which sounds are behaviorally relevant for optimal performance in getting the best outcomes.

Primary motor cortex (M1) is located in the gyrus that is anterior to the central sulcus, and has representations of different portions of the body that output neurons send motor signals to; this organization is known as the homunculus[182]. In 1980, Fetz and Cheney showed how M1 neuron firing directly created movements in their respective output muscles with the use of electromyography (EMG) electrodes implanted into the forelimb muscles of rhesus monkeys[183]. Later research would show that M1 encodes information regarding kinematics or movement dynamics[184-186]. Learning and optimizing motor output is critical in order to enhance behavioral performance and receive the most rewards. Motor learning theories postulate that rapid, error-based learning is mediated through the cerebellum, while slower, reward-based learning is mediated through the basal ganglia[187]. Further, it has been shown that there are neural correlates of reward in premotor, supplementary motor, and cingulate motor areas that project to M1[188-190]. While rewards can indirectly modulate M1 neural firing, via motor learning in order to optimize

behavioral output[191-193], rewards are also able to directly influence firing rates of neurons in M1[194-197]. In a study where rhesus monkeys were required to use a planar manipulandum to control an on-screen cursor and make center-out reaches to a target, a small portion of M1 neurons showed an increased firing rate at the time a reward was supposed to occur[195]. Some of this subset of M1 neurons increased their firing rate at the time of reward when reward was delivered, but more of them showed an increase in firing rate when reward was not delivered, due to an error in their reach. As previously mentioned, midbrain dopamine neurons code reward prediction errors, which instructs regions that dopamine neurons innervate which behaviors to continue and which to stop – including M1, which has been shown to have dopamine innervation[198-201]. Parkinson's Disease is characterized by degeneration of midbrain dopamine neurons, and one impairment as a result is an diminished ability to learn new motor skills[202]. In addition to this impairment seen in Parkinson's Disease, it has also been shown that dopamine release in M1 is necessary for learning new motor skills[203,204]. This research adds to the evidence supporting the necessity of rewards, and particularly dopamine modulation, throughout the brain in order to optimize behavior.

All of these 'primary' regions have relatively straight-forward schemas. V1 and A1 have specific input organizations, the cortical columns and tonotopic mapping, respectively, and M1 output is specifically organized to represent different regions of the body. The fact that all of these regions are modulated by rewards speaks to their importance, and that it is essential to have information about rewards represented ubiquitously in the brain for humans and animals to learn to behave in ways that give them the best outcomes.

Regions of the brain that are critically important for other higher-level functions are also modulated by reward. The hippocampus and entorhinal cortex are regions primarily known for memory, representing our self-location and orientation in an environment[205-209]. However, there

are studies that have shown reward can modulate the coding of neurons in these regions. Rats performing a spatial memory task showed that previously rewarded locations enhanced positional decoding in neurons of the medial entorhinal cortex[210]. In addition, fan cells in layer 2a of the lateral entorhinal cortex were shown to group previously rewarded stimuli with new stimuli that are rewarded following the same association schemes[211]. Further, the authors showed that this change in the cognitive map of task-relevant variables was mediated by midbrain dopamine release in the entorhinal cortex, effectively integrating the new memories of cue-associations with existing ones. The hippocampus is critical to spatial navigation and memory, but has also been shown to be modulated by reward[212]. A recent study showed the hippocampus has a specialized cell population that was only active in rewarded areas of the environment they were navigating[213]. The intersection of memory, spatial awareness, and rewards is logical because, again, part of our ability to receive the best rewards relies on remembering successful places, behaviors, and how to perform.

The key takeaway from this is that rewards are important for neural coding all over the brain, and that uncertainty is a key influence on subjective value calculations. From basic sensory regions in the cortex that experience receptive field adaptations to rewards, to the most critical regions in reward-based decision making that guide our ability to learn and optimize our outcomes. Furthermore, midbrain dopamine neurons code reward prediction errors based on subjective value, and have widespread projections that receive graded signals indicating which outcomes were better or worse than expected. The following chapters will outline the impact of uncertainty on behavior and learning, and how midbrain dopamine neuron coding is a critical mediator in uncertain environments. Finally, we will consider how integration of non-human primate genomic

sequencing techniques with behavioral neurophysiology can assist in the determination of cell type-specific and circuit-specific roles in reward-based decision making and learning.

2.0 Rare Rewards Amplify Learning and Dopamine Responses[1]

Making accurate predictions is evolutionarily adaptive. Accurate predictions enable individuals to be in the right place at the right time, choose the best options, and efficiently scale the vigor of responses. Dopamine neurons are crucial for building accurate reward predictions. Phasic dopamine responses code for reward prediction errors: the differences between the values of received and predicted rewards[41,136,214-219]. These reward prediction error responses guide the direction and magnitude of reward learning[220], through associative and extinction learning[44,45]. Likewise, phasic dopamine neuron stimulation during reward delivery increase both the dopamine responses to reward predicting cues and the choices for those same cues[43]. These learning signals are approximated by reinforcement learning algorithms, including Temporal Difference (TD) and Rescorla-Wagner learning models[221,222]. According to standard TD learning, 'reward predictions' are simply point estimates – formally, the temporally discounted sum of future outcomes[222]. The magnitude of these predictions, often determined by the average value of past outcomes, accurately describe the activity of dopamine neurons in well-controlled laboratory settings[221]. However, point estimate predictions reflect neither predicted uncertainty, nor the shapes of reward distributions, and they are not adequate descriptors of behavior[223-226]. Consider that, learning takes longer when rewards are sampled from broader distributions, compared to when they are sampled from narrower distributions[224,225]. Likewise, decision-makers take longer to choose between options when value differences are small, compared to when differences are large[226,227]. These results demonstrate that probability distributions over reward values, and not simply point estimates such

[1] *The contents of this chapter were previously published (Rothenhoefer et al., 2021).*

as the mean, influence learning and decision making. Dopamine responses adapt to the range or standard deviation of predicted outcomes[42], but it remains unknown if the weights allocated to the tails of reward distributions – a parameter that determines distribution shape and frequency of prediction errors – affects dopamine responses and neural learning rules.

Reinforcement Learning (RL) has produced remarkable advances in artificial intelligence[228,229] and RL techniques have recently been extended to learning probability distributions[230]. Distributional RL models simultaneously learn different value predictions that, together, represent probability distributions. It was recently shown that a range of value predictions derived from distributional RL were reflected by dopamine neurons, raising the enticing possibility that brains employ a distributional code for value[231]. Critically, this distributional code operates at the level of populations, rather than individual neurons. Thus, it is unknown how single dopamine neurons may adapt their responses to predicted reward probability distributions.

To investigate whether the distribution shape differentially affected reward, we created two discrete reward size distributions that reflected, roughly, the shapes of normal and uniform distributions. We trained NHPs to predict rewards drawn from these distributions. Crucially, according to temporal difference (TD) learning, the Normal distribution resulted in rare prediction errors following rewards drawn from the tails of that distribution, whereas the same rewards, with identical prediction errors, were drawn with greater frequency from the Uniform distribution. We found that NHPs learned to choose the better option within fewer trials when rewards were drawn from Normal distributions, compared to when rewards were drawn from Uniform distributions. Moreover, we found that pupil diameter was enhanced with learning from rare prediction errors, but not with learning from common prediction errors of the same magnitude. This result suggests greater vigilance to rare outcomes. Using single neuron recording, we show that dopamine

responses reflect the shape of predicted reward distribution. Specifically, rare prediction errors evoked significantly larger responses than common prediction errors with identical magnitudes. These results demonstrate a complementary but distinct mechanism from TD-like reward prediction error responses for learning based on probability distributions. Due to the identical expected values, but varying distributions of outcomes, these results demonstrate a complementary but distinct mechanism compared to TD-like reward prediction errors for learning based on probability distributions.

2.1 **Methods**

2.1.1 **Animals, Surgery and Setup**

All animal procedures were approved by Institutional Animal Care and Use Committee of the University of Pittsburgh. We used two male Rhesus macaque monkeys (*Macaca mulatta*) for these studies (13.9 and 11.2 kg). A titanium head holder (Gray Matter Research) and a recording chamber (Crist Instruments, custom made) were aseptically implanted under general anesthesia before the experiment. The recording chamber for vertical electrode entry was centered 8 mm anterior to the interaural line (Fig. 5). During experiments, animals sat in a primate chair (Crist Instruments) positioned 30 cm from a computer monitor. During behavioral training, testing and neuronal recording, eye position was monitored using infrared eye tracking (Eyelink Plus 1000). Licking was monitored with an infrared optical sensor positioned in front of the juice spout (Balluff). Eye, lick and digital task events were sampled at 2 kHz. Custom-made software (Matlab, Mathworks Inc.) running on a Microsoft Windows 7 computer controlled the behavioral tasks.

2.1.2 **Behavioral Tasks**

Pavlovian Task

Three distinct cues (fractal images) were used to predict reward. One predicted a sure reward of 0.4 ml. Another predicted a uniform distribution, where 0.2, 0.4, and 0.6 ml were delivered with equal frequency (1/3 probability for each reward). A final cue predicted a normal reward distribution, where 0.2 and 0.6 ml were delivered with low frequency (2/15 probability for each of the two rewards), and the middle reward (0.4 ml), was delivered with a much higher frequency (11/15 probability). Finally, there was an unpredicted reward condition, where 0.4 ml of juice would be given after no cue was presented. In each trial, one of the three cues, or no cue, was pseudorandomly chosen and was presented to the animal. The reward was delivered 2 s after the cue onset. At the same time, and to assist in reward size identification, a value bar cue was displayed on the screen that indicated the reward volume they received. Trials were separated with inter-trial intervals of 2-5 s, chosen from a truncated exponential distribution. Before recording, all cues were well learned after experiencing them repeatedly over multiple sessions (Monkey B: 10 sessions, ~2800 trials; Monkey S: 6 sessions. ~2600 trials).

Choice Task for Measuring Distributions Values

For the data presented in Figure 4b-d, three cues (Fig. 4a) predicted a Normal distribution (Fig. 8a, right), and three different cues (Fig. 4a) predicted a Uniform distribution (Fig. 8a, left). One small, ‘safe’ cue predicted 0.2 ml of juice and one large, ‘safe’ cue predicted 0.6 ml of juice (Fig. 4a). Monkey S was offered binary choices between Normal and Uniform distribution-predicting cues, and between distribution-predicting cues and safe cues. Following successful

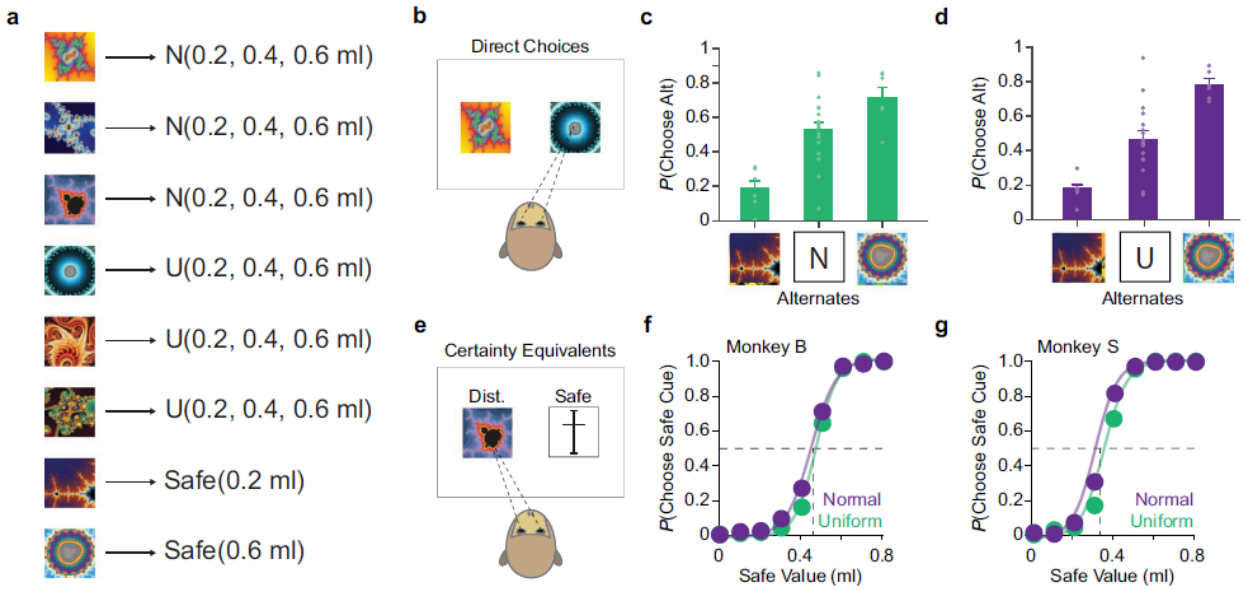


Figure 4: Normal and uniform reward size distributions have equivalent subjective values

a, Schematic of the distribution-predicting fractal cues used to represent Normal (N) and Uniform (U) distributions, and safe values for the choice task in **b**. Three unique cues were used to predict a Normal distribution of rewards, and three unique cues were used to predict a Uniform distribution of rewards. All the distribution predicting cues were comprised of the same three reward volumes (0.2, 0.4, and 0.6 ml), and thus the same expected value (EV) of 0.4 ml. Additionally, one fractal cue predicted a sure reward of 0.2 ml, and another fractal cue predicted a sure reward of 0.6 ml. **b,** Monkeys made saccade-guided choices between Normal distribution-predicting cues, Uniform distribution-predicting cues, and safe rewards. **c,** Bar graphs are the probability of choosing the alternate cue over a Uniform distribution-predicting cue with an EV of 0.4 ml. The alternates from left to right on the x axis are a safe cue predicting 0.2 ml, a Normal distribution-predicting cue with a mean of 0.4 ml, and a safe cue predicting 0.6 ml. Data points are from individual blocks, and error bars represent ±s.e.m. across blocks (between 6 and 18 blocks per condition). **d,** Same as in **c**, but the probability of choosing an alternate cue over a Normal distribution-predicting cue with an EV of 0.4 ml, and the middle alternate option represents Uniform distribution-predicting cues with an EV of 0.4 ml. **e,** The choice task used to measure subjective value. Animals made saccade-directed choices between a distribution predicting cue and a safe alternative option. The safe alternative option was a value bar with a minimum and maximum of 0 and 0.8 ml at the bottom and top, respectively. The intersection between the horizontal bar and the scale indicated the volume of juice that would be received if monkeys selected the safe cue. **f,** Probability of choosing the safe cue as a function of the value of the safe option, when the distribution predicting cue had an expected value (EV) of 0.4ml. Dots show average choice probability for 9 safe value options for monkey B. Solid lines are a logistic fit to the data. Red indicates data from normal distribution blocks, gray indicates data from uniform distribution blocks. The dashed horizontal lines indicate subjective equivalence, and the CE for each distribution type is indicated with the dashed vertical lines. **g,** Same as in **f**, for monkey S.

central fixation for 0.5 s, two choice options appeared on the monitor and the monkey indicated its choice by a saccade towards one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue for 0.5 s to confirm its choice. Reward was delivered 1.5 s later. Trials were separated with inter-trial interval of 1.5-6.5 s, drawn from a truncated exponential distribution. Failure to maintain the central fixation or early break of the fixation on the chosen option resulted in a 4 s time-out, and a repeat of the failed trial.

For the data presented in Figure 4e-g, monkeys made choices between well-learned distribution-predicting fractal cues and 'safe' value bar cues that indicated the magnitude of the alternative option. The value bar cue had a value range of 0 ml to 0.8 ml, in 0.1 ml increments. Wherever the horizontal bar intersected the vertical scale indicated with 100% certainty the size of juice the monkeys would receive if they chose it. The mean of the distribution predicting cues was 0.4 ml. In each choice trial, after successful central fixation for 0.5 s, the two choice options appeared on the monitor and the monkey indicated its choice by a saccade towards one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue for 0.5 s to confirm its choice. Reward was delivered 1.5 s later. Trials were separated with inter-trial interval of 1.5-6.5 s, drawn from a truncated exponential distribution. Failure to maintain the central fixation or early break of the fixation on the chosen option resulted in a 4 s time-out, and a repeat of the failed trial.

Choice Task to Measure Learning

For the data presented in Figure 8, monkeys were offered two never-before-seen cues on the first trial of every block. The block length was selected from a truncated exponential distribution between 15 to 25. Within each block both the cues predicted rewards drawn from the same type of distribution, Normal or Uniform. Further, each novel cue had a different pseudo-randomly selected mean that was either 0.2, 0.3, 0.4, 0.5, or 0.6 ml. For example, if it were a Uniform block, and the means selected for the two cues were 0.3 and 0.6 ml, the rewards for one cue would be 0.2, 0.3, and 0.4 ml (drawn with equal frequency), and 0.5, 0.6, and 0.7 ml (also drawn with equal frequency). In each choice trial, after successful central fixation for 0.5 s, the two choice options appeared on the monitor and the monkey indicated its choice by a saccade

towards one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue for 0.5 s to confirm its choice. Reward was delivered 1.5 s later. Trials were separated with inter-trial interval of 1.5-6.5 s, drawn from a truncated exponential distribution. Failure to maintain the central fixation or early break of the fixation on the chosen option resulted in a 4 s time-out, and a repeat of the failed trial.

Choice Task for Measuring the Subjective Value of Reward Size Distributions

The overall goal of this study was to investigate how predicted distribution shape influenced dopamine responses. To fairly investigate this, we required that the predicted distribution values be the same. Accordingly, we created the Uniform and Normal reward size distributions such that they were composed of the same three elements and had the same Expected Values (Fig. 8a). However, because we planned to record from dopamine neurons and they reflect subjective values, we used two choice tasks to verify that the Expected Utilities (EUs) of the distributions were the same (Fig. 4).

We first used a direct choice task to measure the relative subjective values of the distributions. Visual cues (fractal images) were used to predict rewards. To avoid preferences between cues, we used six different cues to predict distributions – three cues predicted the Normal distribution and three different cues predicted the Uniform distribution (Fig. 4a). To ensure that the monkey was making valid economic choices rather than choosing randomly, we also created two safe cues that predicted a small (0.2 ml) and large (0.6 ml) reward. We reasoned that if subjects were making valid economic choices, they should choose the large reward option over both distributions, and both distributions over the small reward option[232]. We used classical conditioning to train monkeys on the cue-reward contingencies, then we measured binary choices

between the cues (Fig. 4b). The monkey selected the Normal cue over the Uniform cue with a probability of 0.53 ± 0.19; this was not significantly different from chance (Fig. 4c) ($p = 0.48$, $N = 9$ cue pairs, t-test). Additionally, the monkey chose the Normal distribution over the small reward (Fig. 4c, $p < 0.0001$, t-test) and the large reward over the distribution (Fig. 4c, $p = 0.0004$, t-test). Similarly, the monkey chose the Uniform Distribution over the small reward, and the large reward over the distribution (Fig. 4d, $p = 0.001$ and 0.005, respectively $N = 3$ cue pairs, each, t-test) Thus, while making valid economic choices, the monkey was choice indifferent between the distributions. These results provide strong evidence that the predicted values of the two distributions were the same.

The EUs were critical to our interpretation of the data, and as such, we replicated this result using a different behavioral paradigm: we independently measured the certainty equivalents (CEs) of Normal and Uniform reward distributions. CEs are the volumes of rewards the subject would exchange for a gamble; in these experiments the distributions were the gambles. Monkeys made choices between cues that predicted a distribution and cues that explicitly indicated safe options (Fig. 4e, Methods). We plotted the probability of choosing the safe option as a function of the safe option volume and generated psychometric functions (Fig. 4f, g). The CEs were the safe values that corresponded to P(Choose Safe) = 0.5 (black dashed lines in Fig. 4f, g). Analysis of the session-by-session CEs for the Normal and Uniform blocks found no effect of the distribution type on the CEs ($p = 0.2$, $N = 18$. T-test). Therefore, the CEs strongly agree with the direct choice data indicate that the Normal and Uniform reward size distributions had similar subjective values. These results indicated that the prediction errors generated from the distributions could be readily compared and ensured that disparities between prediction error responses were not driven by differences in the predicted subjective values.

2.1.3 **Neuronal Data Acquisition and Analysis of Neuronal Data**

Custom-made, movable, glass-insulated, platinum-plated tungsten microelectrodes were positioned inside a stainless-steel guide cannula and advanced by an oil-driven micromanipulator (Narishige). Action potentials from single neurons were amplified, filtered (band-pass 100 Hz to 3 kHz), and converted into digital pulses when passing an adjustable time–amplitude threshold (Bak Electronics). We stored both analog and digitized data on a computer using custom-made data collection software (Matlab).

Dopamine neurons were functionally localized with respect to (a) the trigeminal somatosensory thalamus explored in awake monkeys (very small perioral and intraoral receptive fields, high proportion of tonic responses, 2-3 mm dorsoventral extent[233], (b) tonically active position coding ocular motor neurons and (c) phasically direction coding ocular premotor neurons in awake monkeys. Individual dopamine neurons were identified using established criteria of long waveform (> 2.5 ms, Fig. 5a) and low baseline firing (< 8 impulses/s)[234]. Following standard sample sizes used in studies investigating neuronal responses in non-human primates, we recorded extracellular activity from 67 dopamine neurons. Forty neurons had a sufficient number of trials and we used these neurons for further analysis.

The neurons that met these criteria showed the typical phasic activation after unexpected reward (Fig. 5b, $p < 0.0001$, $N = 40$ neurons; Wilcoxon rank-sum test). Figure 5c and d show maps of our recording locations relative to both monkeys' grids, and the number of cells recorded at each location. Figure 5e and f show MRI images of monkey S and the location of the recordings.

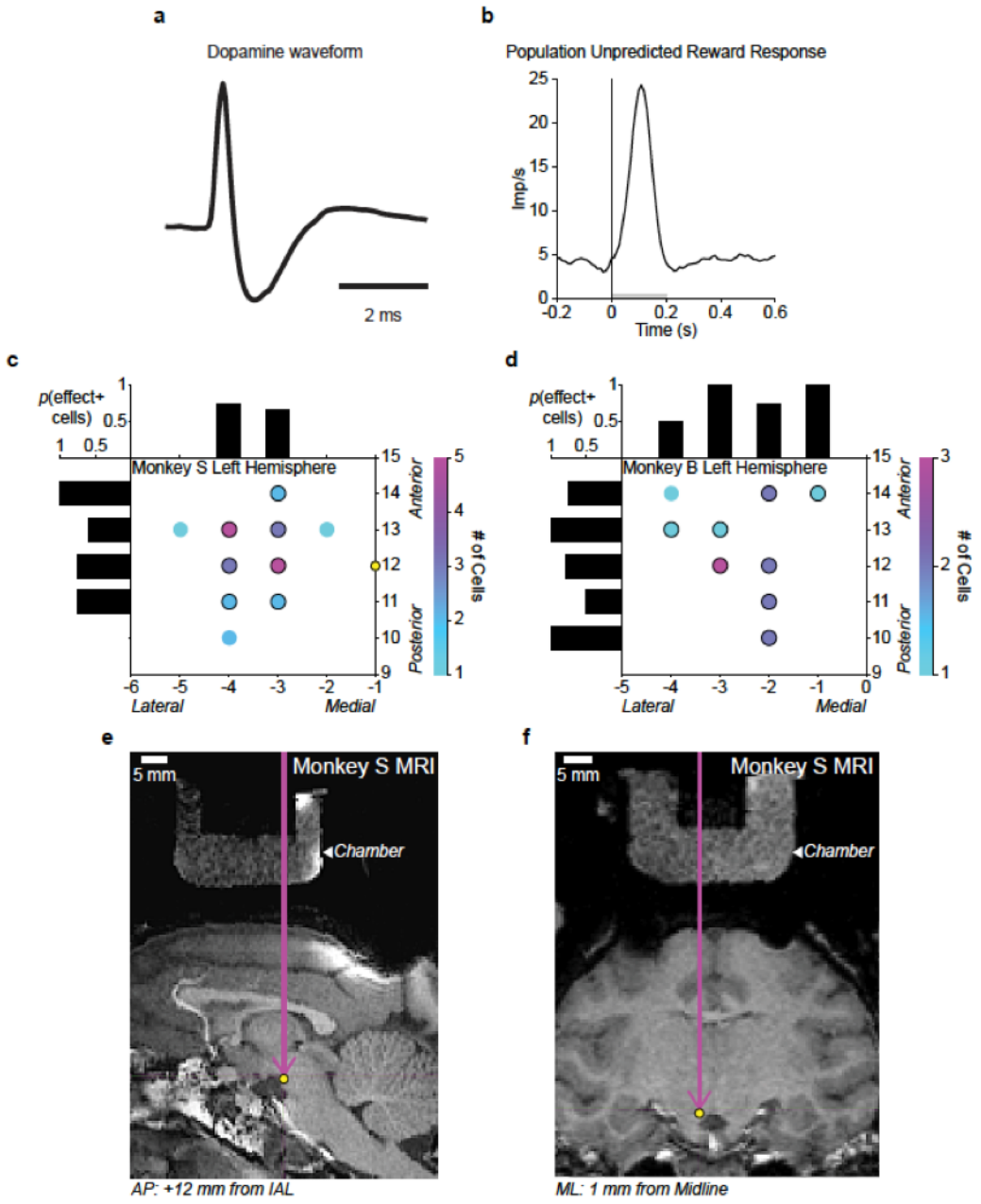


Figure 5: Dopamine neurons and recording sites

a, Example dopamine waveform from one of the neurons in our population. **b,** The population of 40 neurons used for our analyses in the Pavlovian and choice task had significant activations following unpredicted rewards – a characteristic feature of dopamine neurons. Gray bar along the x axis indicate the response window used for analysis. **c,** Recording locations for the left hemisphere of monkey S. X axis indicates lateral to medial location in the grid in millimeters, relative to midline (0). Right y axis indicates posterior to anterior location in the grid in millimeters, relative to interaural line (IAL). Each locations' color indicates the number of neurons recorded for that location. Black circles surrounding the individual locations indicated that neurons recorded there were part of the population of 29 neurons that had steeper response slopes in normal compared to uniform condition. Bar graphs on the left and top axes indicate the proportion of cells in that AP (left) or ML (top) location that were effect positive. Yellow dot corresponds to location indicated in MRI scan shown in d and e. **d,** Recording locations for the left hemisphere of monkey B. Same as panel c. **e,** Sagittal view MRI of the recording chamber of monkey S. Purple arrow indicates the AP location in the grid (+12 mm from IAL). **f,** Coronal view MRI of the recording chamber of monkey S. Purple arrow indicates the ML location in the grid (1 mm from Midline). Yellow dot in e and f correspond to approximate recording grid location in **c**.

2.1.4 **Analysis of Behavioral Data**

Logistic regression

We used logistic regression to quantify the influence of reward distribution on monkeys' behaviors, controlling for trial numbers since a new block starts and the difference between the values of two cues.

$$\log\left(\frac{P(Correct)}{1-P(Correct)}\right) = \beta_0 + \beta_D * D + \beta_C * C + \beta_T * T$$

(Eq. 3)

where *D* is a binary variable for reward distribution type (0 for Uniform and 1 for Normal), *C* is a continuous variable for the difference between the values of two cues and *T* is a categorical variable for the trial number since the start of a new block.

Reinforcement Learning Model

We used a fixed learning-rate reinforcement learning (RL) model to examine monkeys' choices during learning and to acquire trial-by-trial estimate of chosen and unchosen values [222]. The model had two value functions representing the learned values of probability distribution 1 ($pd1$) and probability distribution 2 ($pd2$) respectively. In each trial (t), the probability that the model chooses $pd1$ over $pd2$ was estimated by the softmax rule as follows:

$$P(pd1)_t = \frac{e^{V_t(pd1)/\beta}}{e^{V_t(pd1)/\beta} + e^{V_t(pd2)/\beta}}$$

(Eq. 4)

where β, the temperature parameter of the softmax rule, determines the level of choice randomness.

In each trial, upon making a choice and receiving an outcome, the value of the chosen option on that trial, V_t, was updated according the reward prediction error, as follows:

$$V_{t+1} = V_t + \alpha * \delta_t$$

(Eq. 5)

where α denotes the learning rate, and the prediction error is calculated as the following: $\delta_t = r_t - V_t$, indicates the difference between the predicted and realized reward sizes, V_t and r_t, respectively. The free parameters, α and β, were fit by maximizing the likelihood of the model. After fitting the model, we took the trial-wise mean of the unsigned PE over blocks of the same type (Fig. 8e).

To characterize the transition from active learning to asymptotic behavior, we fit logarithmic functions to each block, and the collected the block-by-block transition trials that marked the crossing of a predetermined threshold that separated active learning from asymptotic behavior. When the first derivative of the fitted prediction errors decreased below a predetermined threshold, we considered that the animal had stopped actively learning. When the magnitude of the prediction errors stayed below 0.1 for more than two trials, we considered that the animal successfully estimated the true value, since the true difference between the lowest/highest values from the mean was 0.1 ml. We designated the boundary between active learning and asymptotic phases as the trial when both conditions were met. The faster learning exhibited in the Normal distribution block was robust under a wide range of prescribed thresholds.

2.1.5 **Analysis of Neural Data**

Data Pre-Processing

We constructed peri-stimulus time histograms (PSTHs) by aligning the neuronal impulses to task events and then averaging across multiple trials. We smoothed the PSTHs by convolving with $(1 - e^{-t})e^{-t/T}$, where T is set to be 20 ms. The analysis of neuronal data used defined time windows, individual to each neuron, that included the major positive and negative response components following cue onset and juice delivery, as detailed for each analysis and each figure caption. The neural activity within time window following juice delivery was baseline-corrected by subtracting the average activity from -1000 ms to 0 ms relative to cue onset.

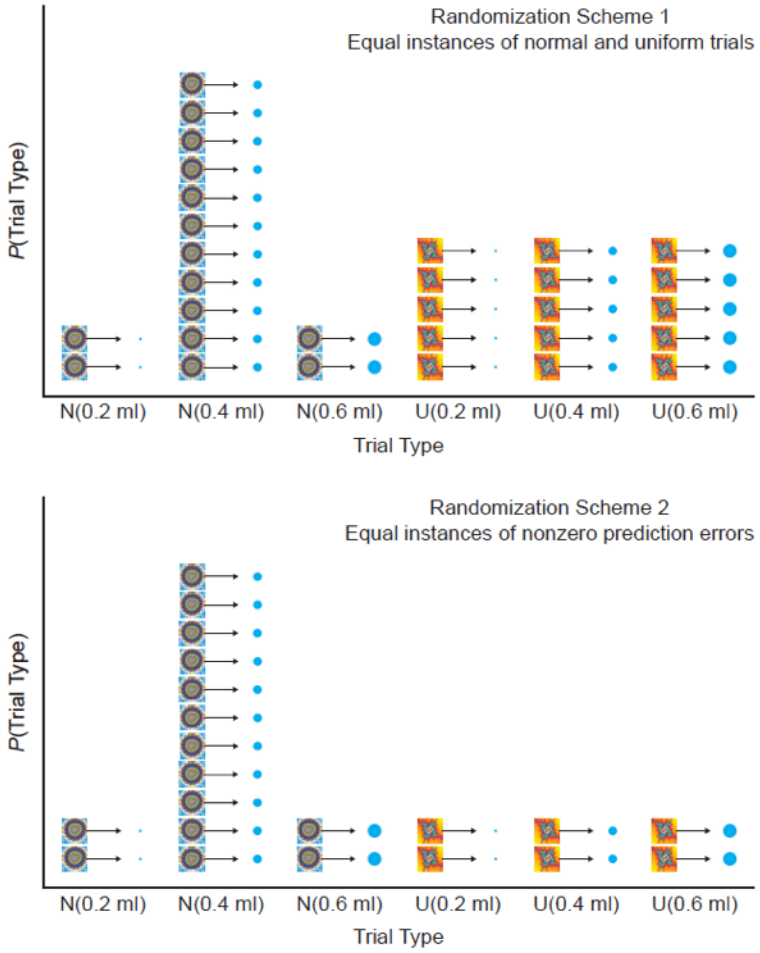


Figure 6: Reward randomization schemes used to determine trial types

Top, 'CS matched" randomization with equal frequencies of Normal and Uniform trials. **Bottom**, "PE matched" randomization with equal frequencies of 0.2 ml and 0.6 ml reward trials in each distribution. In both graphs, the y axis represents the probability of drawing the trial type (trial types drawn with replacement). The 6 trial types divided according to distribution type (N and U) and reward size (0.2, 0.4 and 0.6 ml). The number of instances in each trial type "stack" indicates the probability of drawing the trial type.

Single Neuron Linear Regression

To determine whether previous rewards influence the current CS response, we fit a linear model to each neurons' CS response, using the rewards from the previous 5 trials as the independent variables. We found that previous outcomes up to 5 trials back did not influence CS response. This result is not particularly surprising in the Normal distribution trials, as the previous 5 outcomes were most often 0.4 ml. This reward magnitude evoked no reward prediction errors. The Uniform distribution, on the other hand, did generate more prediction errors. The lack of a clear learning effect in the Uniform distribution has two main causes, we think. First, trial types were determined at random (Fig. 6). Thus, the previous Uniform trial could be several trials back. Second, the monkeys had experienced the cues so often that the learning rate was likely very low.

To assess if reward responses for an individual neuron were enhanced bidirectionally by rare prediction errors, we fit the following linear model to each neuron:

$$F_z = \beta_0 + \beta_1 * D + \beta_2 * R + \beta_3 * D\ x\ R \quad \text{(Eq. 6)}$$

where F_z is the normalized firing rate in the time window following juice delivery, D is a binary variable for reward distribution type (Normal distribution as reference group), R is a continuous variable for reward magnitude and $D\ x\ R$ represents the interaction effect between reward distribution and reward magnitude. Figure 9f was obtained by scatter plotting each neuron's slope for the Normal distribution against its slope for the Uniform distribution. A paired *t*-test was used to see if the slopes were significantly biased towards Normal distribution.

Decoding Distribution Type

For each of the three reward magnitudes, we used a Gaussian naïve Bayes classifier to decode the Normal and Uniform reward distributions from the average firing rate in the time window following juice delivery[235]. We then used leave-one-out cross-validation to assess the performance of the decoder. The resulting confusion matrix was normalized by the number of trials. After cross-validation, permutation tests with 5000 iterations were performed to see if the accuracy of the decoder is significantly different from chance for each reward magnitude. A decoder including all 40 neurons was not able to correctly classify distribution types above chance. Therefore, we used a Selectivity Index (SI) to select neurons for decoding. The single-neuron SI for a particular reward magnitude was defined as the difference between mean reward responses in two reward distribution, divided by the pooled variance of two conditions.

$$SI = \frac{\overline{F_N} - \overline{F_U}}{\sigma_P}$$

(Eq. 7)

The subset of 11 neurons with the largest SI successfully decoded the predicted distribution from the responses to 0.2 and 0.6 ml (Fig. 9g). To ensure that the rest of the neurons did not encode an opposite effect, we built a classifier with the rest of the neurons (29/40) and did not observe above-chance performance (p = 0.515, p = 0.329, p = 0.549, for 0.2, 0.4 and 0.6 ml respectively, Permutation test).

Reversal-Point Correction

To account for variability of reversal point reported in the literature[231], we corrected the reward response of each neuron by subtracting the estimated reward response of its reversal point. We estimated neuron- and distribution-specific reversal point by splitting the distribution of

responses for each neuron, in each distribution, into two groups. One group contained the trials with activations, and the other group contained the trials with suppressions. We then averaged the reward sizes that were associated with the responses in the two groups, and the reversal points were obtained by taking the mean of the two averages (Fig. 10a). The neural activity corresponding to the reversal point was estimated by plugging the reversal point into the single neuron linear regression described above. For each neuron in each distribution, we subtracted this estimated activity from the responses to 0.2, 0.4 and 0.6 ml. We used a two-tailed Wilcoxon signed rank test to test if neurons with steeper response slopes to rewards from Normal distributions show bidirectional stretch in their reward responses, after reversal point correction (Fig. 10b).

2.1.6 **Deconvolution**

Event-related pupil responses were analyzed trial-by-trial using nideconv[236,237], a Python package that specializes in fMRI and pupil signal deconvolution. The design matrix for a trial consisted of a total of four event types: the onset of central dot for fixation, the onset of cue presentation, the monkeys' saccades to indicate choice, offset of cue presentation (in temporal order), and the onset of reward. The pupil diameter changes related to fixation and the offset of cue presentation were analyzed 0.5 s pre-event until 2 s post-event; the time windows for the onset of cue presentation and monkeys' saccades started 0.5 s pre-event and ended 3 s post-event; the time window for the presence of rewards started at 0.5 s pre-event and ended at 1.5s post-event. To understand the relationship between pupil diameter and prediction error post-reward, reward prediction errors and value estimates derived from the model were used as covariates in the deconvolution algorithm. Consequently, we obtained a measure of how sensitive the post-reward pupil diameter changes are to the prediction errors in each reward distribution, by looking at the

beta coefficients in the prescribed time window. Finally, we grouped the deconvolved signal based on the Active/Asymptotic learning period distinction and reward distribution type and calculated the ensemble average across trials.

2.2 **Results**

2.2.1 **Reward Size Distributions**

We used non-informative images (fractals) to predict rewards drawn from differently shaped distributions. Distribution shapes were defined according to relative reward frequency. One fractal image predicted an equal probability of receiving a small, medium, or large volume of juice reward. We define this as the uniform reward size distribution (Fig. 8a left). A second fractal image predicted that the small and the large reward would be given 2 out of 15 times, and the medium sized reward would be given 11 out of 15 times (Normal reward size distribution, Fig. 8a right). Importantly, both reward size distributions were symmetrical and were comprised of the same three reward magnitudes. Therefore, the Uniform and Normal distributions had identical Pascalian expected values. However, rewards drawn from the tails of the normal distribution were rare, compared to the frequency of identical rewards drawn from the tails of the uniform distribution. Anticipatory licking reflected the expected value of both the distributions, as well as the expected value of safe cues (Fig. 7, p = 0.019, Linear Regression). Thus, the animals learned that the cues predicted rewards.

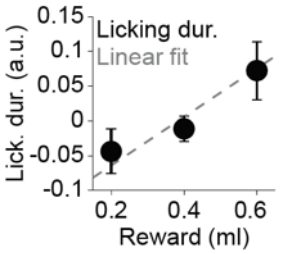


Figure 7: Anticipatory licking show that monkeys learned the predicted reward magnitude

Black dots indicate the normalized licking duration data for predicted reward volume, and the grey dotted line indicates the linear fit to the data. Error bars indicate ±s.e.m. across session.

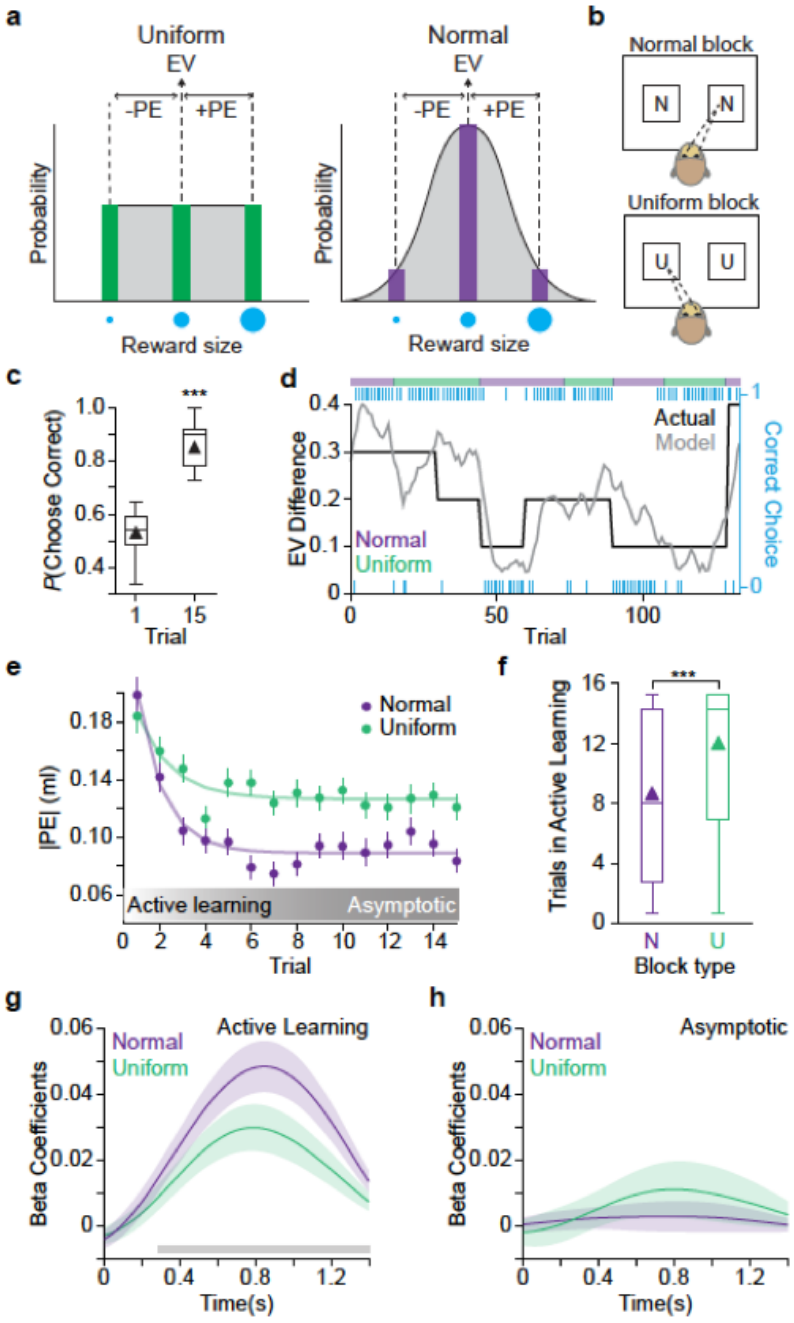


Figure 8: Behavior during distribution choice task

a, Schematics of uniform and normal reward size probability distributions. Green and purple bars indicate the probabilities of different reward sizes in the uniform and normal conditions, respectively. Small, medium and large reward sizes are indicated by small, medium and large blue dots. Gray shaded regions show the underlying probability density functions. EV, expected value; −PE/+PE, negative/ positive prediction error. A cue associated with a uniform distribution predicted that each reward size would be drawn with 1/3 probability. A different cue associated with a normal distribution predicted that the small and large reward volumes would be drawn 2/15 times, whereas the medium reward size would be drawn 11/15 times. **b,** Task where each block was either a normal or a uniform. In normal blocks, two new cues represented normal distributions with different EVs. In uniform blocks, two new cues represented uniform distributions with different EVs. **c,** Box and whisker plots show the probability of choosing the higher-valued option on trials 1 and 15, across both distribution types. Triangles represent averages. ***$p < 0.0001$, N = 275 blocks, t-test. **d,** Reinforcement learning model performance for a subset of trials. Actual (black) and estimated (gray) value differences for two choice options. The bar at the top indicates either normal or uniform block type. The primary y axis shows the EV difference between the two choice options, and the x axis shows trial number. The blue tick marks correspond to correct and incorrect choices, indicated by the secondary y axis. **e,** Absolute PEs as a function of trial number within normal and uniform blocks. The y axis represents the absolute PE (|PE|) in ml of juice. Error bars are ±s.e.m. across 142 normal blocks and 133 uniform blocks, and solid lines are exponential functions fit to the data. Shaded box schematically describes the transition from 'active learning' to 'asymptotic' behavior, the actual transition trials were determined on a block-wise basis (Methods). **f,** Box and whisker plot shows the number of trials in the active learning phase for normal and uniform distribution blocks. ***$p < 0.0001$, Mann–Whitney U-test. **g,** Beta coefficients from a deconvolution analysis on pupil diameter data for trials in the active learning phase of normal and uniform blocks, aligned to reward delivery at time 0. The gray horizontal bar indicates time points after reward where normal beta coefficients were significantly different from uniform beta coefficients ($p < 0.05$, N = 4,703 trials, cluster-based permutation test). Shaded regions indicate 95% confidence interval over trials. **h,** As in **g**, for trials in the asymptotic phase.

2.2.2 **Distribution Shape Affects Learning**

To investigate whether the distribution shape differentially affects reward learning, we created symmetrical reward size distributions that simulated the shapes of Uniform and Normal distributions (Fig. 8a). Within each block (15-25 trials), monkeys made choices between two never-before-seen cues that predicted Normal or Uniform reward size distributions, as in Fig. 8a, with pseudorandomly chosen EVs (Fig. 8b, Methods). As expected, the monkeys performed at chance levels on trial 1, but quickly learned to choose the better option (Fig. 8c). Logistic regression of the choice behavior indicated both trial-by-trial learning and better overall performance in the Normal blocks (β_{Trial} = 0.110, $p < 0.0001$, β_{Normal} = 0.167, $p = 0.007$, N = 6098 trials, *t*-test). We used a standard Reinforcement Learning (RL) model[222] to quantify the prediction errors generated during learning (Fig. 8d, Eq. 4, Eq. 5). This analysis revealed that behavior in both block types was characterized by an active learning phase when the prediction error magnitudes were diminishing, and a later asymptotic phase when the magnitudes were stable (Fig. 8e). However, the number of trials in the active learning phase was significantly fewer in the Normal blocks compared to the Uniform blocks (Fig. 8f, Methods). Together, these data showed enhanced learning performance in blocks with rewards drawn from Normal distributions.

We hypothesized that the animals would learn faster from reward sizes sampled from the normal distribution compared to the uniform distribution. We used a reinforcement learning (RL) model to quantify the learned values (Methods). Our model, fit to the behavioral choices, performed well at predicting the true values of the two choice options (Fig 8d). To differentiate active learning from stable, asymptotic-like behavior, we fit a logarithmic function to the estimated prediction errors (Fig. 8e). When the change in the log-fitted values went below a robust threshold,

we considered the values to be learned (Methods). Using this approach, we determined the block-wise number of trials needed to learn. We found that, on average, animals needed 4 additional trials to learn in the uniform blocks, compared to the normal blocks (Fig. 8f, $p < 0.001$, Mann-Whitney U test). These data demonstrate that the monkeys learned faster from normal reward distributions, compared to uniform reward distribution. Thus, rare prediction errors have greater behavioral relevance than common prediction errors of the same magnitude.

2.2.3 **Enhanced Autonomic Responses to Rare Rewards**

To investigate autonomic responses to prediction errors, we analyzed the pupil responses during the choice task (Fig. 8b). We used deconvolution to separate the effects of distinct trial events on pupil responses (Methods). The deconvolution analysis indicated that during the active learning phase, pupil diameter responses were more sensitive to rare reward prediction errors than to common reward prediction errors of the same magnitude (Fig. 8g, Methods). The gray horizontal bar in Figure 8g indicates time points after reward where the Normal beta coefficients from the deconvolution analysis that were significantly different from the Uniform beta coefficients ($p < 0.05$, $N = 4{,}703$ trials, cluster-based permutation test). This indicates that greater vigilance or arousal was associated with learning from rare-prediction errors. This effect disappeared during the asymptotic phase (Fig. 8h). Thus, pupil responses indicated greater levels of arousal to rare prediction errors during learning.

2.2.4 **Dopamine Responses to Rare Rewards**

After showing that probability distributions were reflected in learning behavior and autonomic responses, we investigated the neural correlates of probability distributions. To do so, we recorded extracellular dopamine neuron action potentials during a passive viewing task (Fig. 9a, Fig. 5, Methods). Here, the magnitudes of the small, medium, and large rewards were fixed at 0.2, 0.4, and 0.6 ml, respectively (Fig. 9a). Prior choice testing confirmed that Normal and Uniform distributions with these reward size elements had equivalent expected utilities (EUs) (Fig. 4, Methods). As expected from cues that predict the same EUs, Dopamine neurons were similarly activated by the Normal and Uniform distribution predicting cues (Fig. 9b). Thus, the passive viewing task rigorously controlled the magnitudes of conventional prediction errors – defined as received minus predicted reward values.

At the time of reward delivery, dopamine neurons are activated or suppressed by rewards that are better or worse than predicted, respectively. Therefore, we expected dopamine neurons to be activated by delivery of 0.6 ml and suppressed by delivery of 0.2 ml. We used two different randomization schemes to control for the number of times each distribution was presented (CS-matched), or the number of times each prediction error was experienced (PE-matched) (Fig. 6). Under both randomization schemes, the 0.6 ml reward activated a larger dopamine response in Normal distribution trials, compared to dopamine activations following delivery of the same volume reward in Uniform distribution trials (Fig. 9c, d, solid lines).

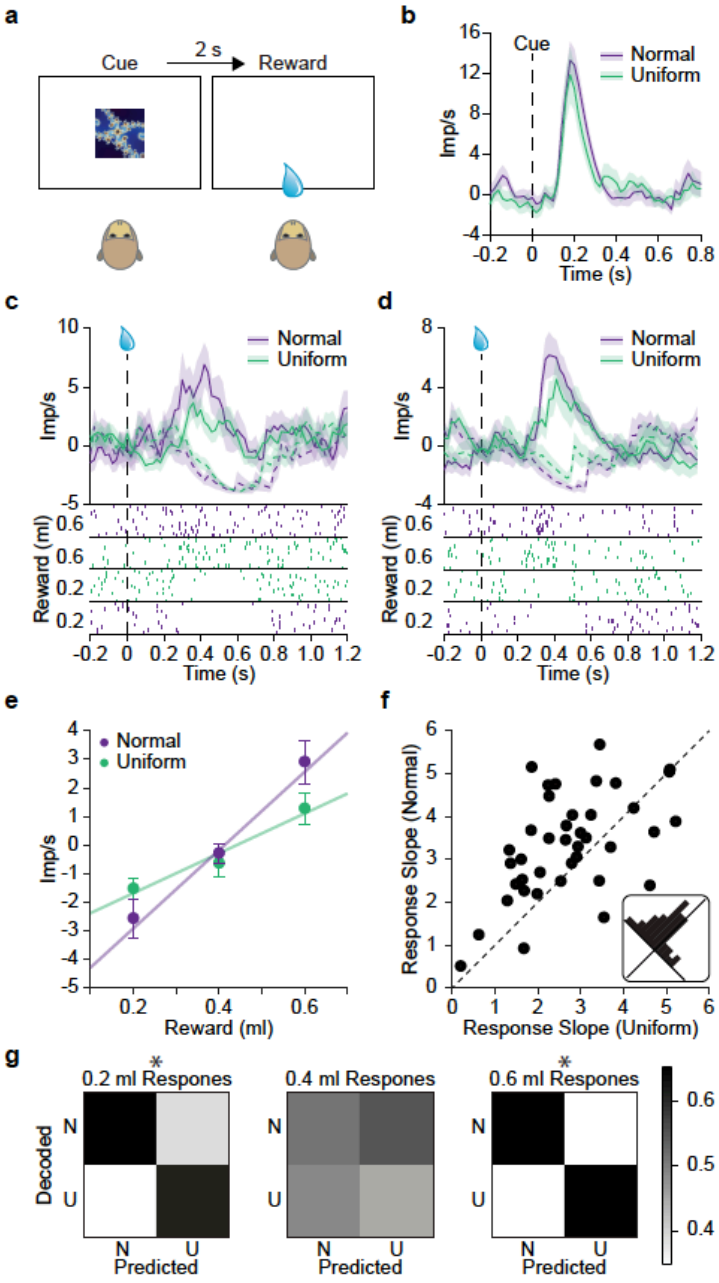


Figure 9: Rare rewards amplified dopamine reward prediction error responses

a, In the recording task, the monkeys viewed a distribution predicting CS and rewards were delivered two seconds later. **b,** Peri-stimulus time histogram (PSTH) of CS-evoked responses to the Normal and Uniform distribution predicting cues in a single neuron. There was no significant difference between the response magnitudes (p = 0.69, N = 40 neurons, Wilcoxon rank-sum). Shaded regions represent ±s.e.m. across trials. **c,** Single neuron reward responses to rewards during Normal and Uniform trials, recorded using the CS-matched randomization scheme (Figure 6). **Top:** PSTHs show impulse rate as a function of time. Solid lines show responses to 0.6 ml, whereas dashed lines show responses to 0.2 ml of juice. Shaded regions represent ±s.e.m. across trials. **Bottom:** Raster plots, separated by Normal and Uniform trials and by reward sizes. Every tick mark represents the time of an action potential, and every row represents a trial. Black vertical dashed line indicates the time of reward. **d,** as in c, for a neuron recorded using the PE-matched randomization scheme (Figure 6). **e,** Single neuron linear regression of a single neuron (c) showed steeper response slopes to rare rewards drawn from the normal distributions. Solid lines indicate the fitted slopes in Normal and Uniform distribution trials. Dots represent the average neural response rewards in Normal and Uniform distribution trials. Error bars represent ±s.e.m. across trials (all data points had between 13 and 76 trials). **f,** Scatter plot of Normal and Uniform distribution response slopes from every neuron (p = 0.003, N = 40 neurons, t = 3.19, t-test). Inset: Histogram shows the density of the dots relative to the diagonal unity line. **g,** Confusion matrices of distribution identity decoding from neuronal responses to 0.2 ml, 0.4ml and 0.6ml rewards in the Normal and Uniform distributions. The matrix sectors are shaded according to the proportion of trials decoded as Normal (N) and Uniform (U). The scale bar on the right shows that darker shades indicate higher proportions. Black asterisks indicate decoding performance above chance level for the responses to 0.2 ml and 0.6 ml (p = 0.045 and p = 0.028, N = 11 neurons, Permutation test, uncorrected p-values). No asterisk above 0.4 responses indicate no significant decoding (p = 0.642, N = 11 neurons, Permutation test).

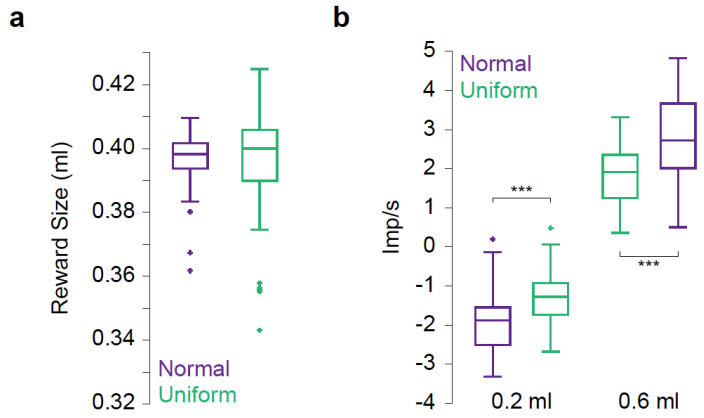


Figure 10: Dopamine pseudo-populations and single neurons simultaneously reflect predicted probability distributions

a, Box and whisker plots show the spread of reversal points for the population of neurons in normal (purple) and uniform (green) trials ((0.0065, 0.0129) and (0.0133, 0.0221), $N = 40$ neurons, bootstrap 90% confidence interval for standard deviation). **b,** Box and whisker plots show the baseline subtracted responses to 0.2 and 0.6 ml of juice. ***$p < 0.0001$, $N = 29$ neurons, Wilcoxon signed-rank test, Bonferroni corrected. Responses to 0.4 ml were not significantly different and so not shown ($p = 0.226$, $N = 29$ neurons, Wilcoxon signed-rank test). Box and whisker plots show the median (line), quartiles (boxes), range (whiskers) and outliers (+).

Likewise, dopamine responses were more strongly suppressed by delivery of 0.2 ml reward during Normal distribution trials, compared to delivery of the same reward during Uniform distribution trials (Fig. 9c, d, dashed lines). Linear regression revealed that thirty-four neurons were significant for reward size, and that the vast majority of neurons (29/40) had steeper slopes for the Normal condition, compared to the Uniform condition (Fig. 9e, f). Thus, rare prediction errors resulted in bidirectional amplification of the responses, compared to common prediction errors of the same magnitude. Moreover, because the amplification was bi-directional – both activations and suppressions were amplified in single neurons – this effect could not be attributed to differences in predicted subjective values. Thus, the effects we observed were robust to errors in the measurement of subjective value. These findings confirm that dopamine responses are sensitive to probability distributions during complex behaviors and demonstrate that amplified dopamine responses can be used to guide active learning and decision making.

We applied a naive Bayes classifier to 11 neurons with the greatest selectivity for rare rewards (Methods). The classifier was able to decode distribution identity from the responses to 0.2 ml and 0.6 ml, but failed to decode the distribution from the responses to 0.4 ml (Fig. 9g).

Together, these results demonstrate that phasic dopamine responses reflect predicted probability distributions.

2.2.5 **Reversal Point Variability in Distribution Predictions in Dopamine Neurons**

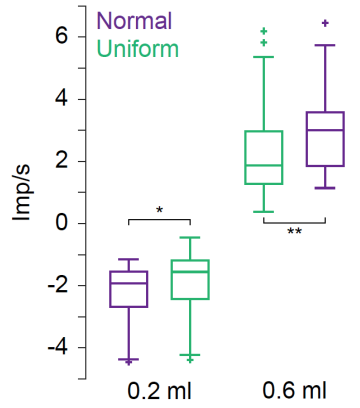


Figure 11: Amplification effect was robust

Box and whisker plots show the baseline subtracted responses to 0.2 and 0.6 ml of juice, as in Fig. 8, but applied to all 34 neurons that were significantly modulated by value. * indicates $p < 0.05$, ** indicates $p < 0.01$, $N = 34$ neurons, Wilcoxon signed-rank test, Bonferroni corrected for multiple comparisons. Box and whisker plots show, median (line), quartiles (boxes), range (whiskers), and outliers (+).

Finally, we investigated whether reversal point variability reflected the predicted distributions. We categorized responses as activations or suppressions and calculated the reversal points for each neuron in each distribution (Methods). As predicted by the distributional TD model[231], the Uniform distribution evoked a larger spread of reversal points compared to the Normal distribution (Fig. 10a). We subtracted cell- and distribution-specific reversal points from each cells' average responses to the three different rewards and tested whether the differential reversal points accounted for the bidirectional response amplification. Following reversal point correction, we still observed significantly amplified responses, in both the negative and positive domain, to identical rewards drawn from the Normal compared to the Uniform distribution (Fig. 10b, Fig. 11), but no significant difference in the reversal point-corrected responses to 0.4 ml. These results demonstrate that the bidirectional amplification of responses is not accounted for by the reversal points. Moreover, these results hint that the single cell-level amplification of responses and the population level distributional TD model could be complementary schemes for learning the shapes of probability distributions.

2.3 **Conclusion**

We dissociated unpredictability from expected value by pseudorandomly drawing rewards from symmetric probability distributions with equal values. These data show that rare prediction errors – compared to commonly occurring prediction errors of the same magnitude – evoke faster learning, increased autonomic arousal, and enhanced dopamine reward prediction error responses. Monkeys learned faster from the more unpredictable rewards drawn from the tails of normal distributions and showed larger pupil responses when learning from rare rewards. In addition, amplified dopamine responses were evident even when identically sized rare and common rewards generated identical TD prediction errors. This result demonstrates that dopamine responses are sensitive to the shapes of predicted probability distribution, rather than just the predicted mean.

Our data show that dopamine-dependent learning behavior and dopamine reward prediction error responses reflect probability distributions. Together, these data reveal a novel computational paradigm for phasic dopamine responses that is distinct from, but complementary to, conventional reward prediction errors. Rare events are often highly significant[238], and our data show decision-makers exhibit greater vigilance towards rare rewards and learn more from rare reward prediction errors. Amplified dopamine prediction error responses provide a mechanistic account for these behavioral effects.

More than 20 years ago, single unit recordings demonstrated the importance of unpredictability for dopamine neuron responses[215]. Since then, the successful application of TD learning theory to dopamine signals has largely subsumed the role of unpredictability, and recast it within the framework of value-based prediction errors[35,239]. In most experimental paradigms, reward unpredictability is captured by frequentist probability of rewards and, therefore,

unpredictability is factored directly into expected value[218,219,240-242]. The resulting Pascalian expected values are learned by TD models[222] and reflected in dopamine responses[218,240,242]. Here, we dissociated unpredictability from expected value by pseudorandomly drawing rewards from symmetric probability distributions with equal values. Monkeys learned faster from the more unpredictable rewards drawn from the tails of normal distributions. Likewise, dopamine responses were amplified by greater unpredictability, even when the conventional prediction error was identical. These data reinforce the importance of unpredictability for dopamine responses and learning.

Several lines of evidence indicate that the amplifications of dopamine responses were not explained by differences in conventional prediction errors. Behavioral assays showed that the monkeys assigned similar Expected Utility (EU) to both distributions (Fig. 4). EU is a proxy for the 'predicted reward value' term used to describe dopamine reward prediction error responses[36]. Accordingly, dopamine responses to Normal and Uniform distribution predicting cues were indistinguishable (Fig. 9b). Therefore, the amplified dopamine responses we observed here were not explained by differences in conventionally defined prediction errors. Rather, the dynamic ranges of the neurons adapted to the shapes of the predicted probability distributions (Fig. 9c-f).

Biological learning signals have inspired deep reinforcement learning algorithms with performance that exceeds expert human performance on Atari games, chess and Go[228,243]. Recently, a new machine learning model, distributional RL, was applied to study the activity of dopamine neurons[231]. A fundamental distinction between distributional RL and the results we present here is the scale at which outcome distributions are represented. In distributional RL, the probability distribution is represented at the level of dopamine neuron populations. In contrast, our results show that single dopamine neurons are sensitive to the shape of the probability distribution

(Fig. 9c–f). Our data indicate that two mechanisms, one operating at the level of populations and the other at the level of single neurons, are complementary schemes for learning probability distributions. Indeed, our data confirmed one prediction from the distributional RL model: for the same population of dopamine neurons, the spread of the measured reversal points is larger for uniform, when compared to normal predicted reward distributions (Fig. 10a). Nevertheless, even after accounting for the distribution-sensitive reversal points, we still observe bidirectional amplification of dopamine responses to rare rewards (Fig. 10b). These results reveal complementary learning schemes within the same population of dopamine neurons.

At the level of single neurons, the amplified dopamine responses to rare rewards indicate that reinforcement learning (RL) models that acquire only point estimate predictions are not adequate to describe dopamine activity. Rather, these data suggest that RL algorithms that track uncertainty, such as Kalman TD[232], may provide an appropriate conceptual framework to explain information processing in the reward system. Kalman-like reinforcement signals enables reward prediction and estimation of uncertainty[244], and therefore may be critical for implementing Bayesian inference. In this sense, the observed amplification of dopamine responses by rare rewards is consistent with a signal that could guide Bayesian inference of the most likely outcomes. Nevertheless, future studies will be required to understand whether phasic dopamine responses can support explicit Bayesian inference for optimal economic choices.

The amplification of dopamine responses by rare rewards appears to be a distinct phenomenon from novelty driven dopamine responses that we and others have previously observed[240,245]. Stimulus novelty decays with the number of exposures and dopamine responses appear to follow this decay[240]. A recent study has shown that stimulus novelty, specifically, and not rarity, drives the large dopamine responses observed during the first exposures to stimuli, and

that novelty-driven CS responses promote learning[245]. None of the rewards used in our study were novel, as only three rewards were used while recording: 0.2 ml, 0.4 ml, and 0.6 ml of blackcurrant juice. The monkeys experienced these three rewards hundreds of times each during every session. Rarity was maintained only during Normal trials, when 0.2 ml and 0.6 ml were rarely given. Thus, the amplification of dopamine responses to rewards drawn from the tails of Normal distributions is likely a function of reward rarity, and distinct from novelty responses.

Pupil diameter was sensitive to prediction errors generated during active learning phases, but the sensitivity sharply decreased after learning. This result is consistent with prior studies showing that pupil diameter is more sensitive to unexpected uncertainty, compared to expected uncertainty[246], and indicates that the monkeys learned the distributions and their associated expected uncertainties. In parallel, we observed that learning was enhanced in the Normal distribution trials. Specifically, we observed that learning became asymptotic after fewer trials in Normal compared to Uniform blocks. Together, these results are consistent with prior studies in humans showing that reward learning is dependent on standard deviation and higher statistical moments of reward distributions[223-225]. Further experiments will be required to disentangle the effects of higher statistical moments, especially standard deviation and kurtosis, on reward learning.

One limitation of our study is that the behavioral choice data and these neural recordings were collected using different tasks. The behavioral paradigm enabled us to directly measure learning differences, however, it required models to do post-hoc estimation of the underlying reward prediction errors. This dependency on model-derived estimates constrained our ability to control the magnitudes of reward prediction errors. Therefore, we used a passive-viewing task to control prediction errors during neuronal recordings (Fig. 9a). This strategy of measuring behavior

in one version of the task and doing neural recordings in a simplified version of the task has been used many times previously by ourselves and others[247]. However, the experimental separation of the behavioral measurement from the neural recordings prevents us from drawing firm conclusions regarding the role of dopamine signal amplifications in learning. Future studies that combine complex behavior and neural recording in the same task will be critical for determining the trial-by-trial relationship between dopamine response amplification and behavior.

It is tantalizing to speculate about the possibility that the neural circuits responsible for value processing evolved in a world where the Normal distribution makes frequent appearances – and that this evolutionary history makes it easier for individuals (and their dopamine neurons) to learn Normal statistics. Regardless, the amplified dopamine responses coupled with the faster learning dynamics observed here suggest that the magnitude of dopamine release may affect cellular learning mechanisms in the striatum. Moreover, dopamine responses have the ability to modulate dopamine concentrations in the prefrontal cortex (PFC), which are tightly linked to neuronal signaling and working memory performance[248]. These findings raise the possibility that amplified dopamine responses could contribute to the exaggerated salience of rare events and postulate a neural mechanism to explain aberrant learning behaviors associated with debilitating mental health disorders such as psychosis, schizophrenia, and depression.

3.0 Ambiguity Preferences Depend on Reward Magnitude

Decision theory recognizes two forms of uncertainty: risk – in which the underlying probability distributions are known, and ambiguity – in which the underlying probability distributions are not known[4]. Expected Utility Theory (EUT), is the dominant economic theory of decision making, describes decisions under uncertainty[5]. In this framework, decision-makers combine probability with subjective value, and select the option with the highest expected utility. Non-human primates (NHPs) and humans incorporate risk and reward magnitude to generate choice behavior that is estimated by EUT[249]. Ambiguity aversion, first illustrated in the Ellsberg paradox, is a phenomenon where decision-makers will avoid ambiguity at the cost of utility – violating EUT[1]. This is a well-documented behavioral paradox that is present in humans and non-human primates[82,250-254]. A study that measured decision making in humans showed that learning about ambiguity aversion made them more tolerant to ambiguity, but still did not fully prevent ambiguity aversion[255]. Moreover, decision-makers also became less risk averse, indicating an incorrect generalization to risk from ambiguity. Many decision making theories based on risk, including EUT, fail to adequately describe real-world choice behavior, as illustrated by the Ellsberg paradox. This failure partly reflects the fact that pure risk is rare: it is only encountered in casinos, coin-flips, and decision making experiments. In most real-world cases, incomplete information, sparse data, and cognitive limitations create various states of ambiguity. Therefore, ambiguity, rather than risk, better describes the conditions of uncertainty under which most real-world decision making occurs. Prior work has demonstrated that non-human primates demonstrate ambiguity aversion during choices between risky and ambiguous stimuli[254]. In addition, it has been shown that the magnitudes of rewards NHPs choose between in risky decisions modulates their

risk preference, such that they are risk seeking with small reward gambles and risk averse with large reward gambles[256]. It is currently unknown whether NHPs will show this shift in preference for ambiguity as they do for risk, depending upon whether they are given low- or high-stakes decisions versus safe, certain value options.

Pupil diameter is closely related to attention, vigilance, learning, and has even been shown to reflect belief state[257-259]. Uncertainty in an individual's environment triggers the sympathetic nervous system[260-262]. Activation of the sympathetic nervous system causes norepinephrine release from the locus coeruleus, leading to the physiological "fight-or-flight" responses[263,264]. These physical reactions include increases in heart rate, perspiration, and pupil diameter[265]. Pupil diameter has been shown to be tightly correlated with unexpected uncertainty and errors in expectation, effectively signaling that learning needs to take place to update predictions about the environment[266]. In Chapter 2, we show that during learning, pupil responses to reward prediction errors following risky cues with higher uncertainty are suppressed in comparison to less uncertain risky cues[267]. This suggests that the lack of arousal from norepinephrine-mediated circuits could minimize the speed with which animals learn in risky conditions with higher uncertainty. It is currently unknown whether pupil responses will show differences in response to risky and ambiguous options with the matched levels of uncertainty.

Here, we investigate whether decision-makers will have varying levels of ambiguity preference dependent upon the EV, like they have EV-dependent risk preferences[256]. We developed a novel NHP economic decision making task with independent control over the level of uncertainty, and reward ranges, magnitudes and probabilities. By matching EV, and manipulating only the amount of uncertainty, measured by Shannon entropy[268], we hypothesized we could reveal a specific effect of ambiguity on decision making and pupil diameter. On a subset of trials, the

outcome probability information was fully obscured to create ambiguity in the outcome probability distribution. During ambiguity trials with small reward sizes, we observed ambiguity-seeking tendencies: the animal selected ambiguous options over safe rewards with similar values, and even valued them more than the risky cues of the same EV. With larger rewards, the trend was reversed: the animal was ambiguous-averse, and switched his preference and valued risky options more than the ambiguous one. Thus, similar to what is seen with risk preferences[256], the monkey displayed value-dependent ambiguity attitudes, such that it became ambiguity averse as the EV increased. In addition to the choice data, we analyzed pupil diameter and found that pupil dilation was less sensitive to ambiguous cues, and the reward volumes received following, at both low and high EVs. Pupil diameter has been previously shown to reflect learning rate and prediction error[266,267]. Thus, the decreased sensitivity to the ambiguous cues and their subsequent rewards indicates that those predictions might not be updated quickly. Together, these observations suggest that there are fundamental distinctions between decision making under risky or ambiguous contexts.

3.1 **Methods**

3.1.1 **Animals, Surgery and Setup**

All animal procedures were approved by Institutional Animal Care and Use Committee of the University of Pittsburgh. We used one male Rhesus macaque monkeys (*Macaca mulatta*) for these studies (11.2 kg). A titanium head holder (Gray Matter Research) was aseptically implanted under general anesthesia before the experiment. During experiments, animals sat in a primate chair (Crist Instruments) positioned 30 cm from a computer monitor. During behavioral training, testing

and neuronal recording, eye position was monitored noninvasively using infrared eye tracking (Eyelink Plus 1000). Eye and digital task event signals were sampled at 2 kHz. Custom-made software (Matlab, Mathworks Inc.) running on a Microsoft Windows 7 computer controlled the behavioral tasks.

3.1.2 **Behavioral Tasks**

Entropy as a Method to Quantify and Compare Ambiguity and Risk

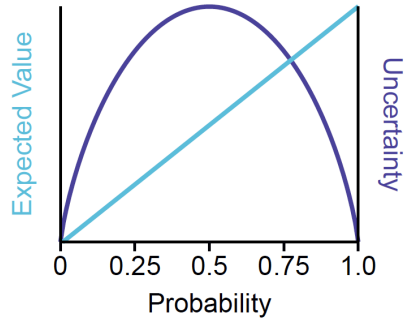


Figure 12: The relationship between expected value and uncertainty

In a two-outcome gamble, where one outcome is to receive reward, and the other is to receive no reward, as probability increases, the effect on expected value (primary axis, light blue) and uncertainty (secondary axis, purple) differ. Specifically, as probability (p) of receiving reward increases in relationship to the probability of not getting reward (1-p), expected value increases linearly. Uncertainty as a function of probability is an inverted U shape. This being the case, the most uncertainty in a two-outcome gamble is when the options are equiprobable, or a 50% probability for each outcome. As one of the two outcomes becomes more probable over the other, uncertainty diminishes due to the increasing likelihood (or certainty) of one of the outcomes. This being the case, there is no uncertainty when the probability of one outcome is 100% and the other is 0%.

In order to compare the effect of ambiguity on behavior and sympathetic responses compared to risk, we used Shannon entropy to quantify the amount of information (or uncertainty) in risky and ambiguous cues[268]. Entropy can be described as the amount of random information, and when using cues with discrete outcomes, entropy is measured in bits. Entropy, or bits of random information, using the following equation:

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_2 \mathrm{P}(x_i) \qquad \text{(Eq. 8)}$$

The function is a sum over all the cues' possible outcomes (n), and their probabilities (x). Given this function, a cue that has a 100% probability of giving one reward volume has 0 bits of random information, and an equiprobable gamble between two different reward magnitudes is equal to 1 bit of random

information. Figure 12 shows the relationship between uncertainty (or entropy, light blue) and expected value (purple) in two outcome gambles, when the probability of getting a reward (p) is between 0 and 1, and the probability of getting no reward has a probability of 1-p. The expected value of the gamble increases linearly as the probability of reward gets higher and the probability of no reward gets lower. Uncertainty is at its peak when the gamble is equiprobable, and at its lowest when there was 0 probability of one of the outcomes and 100% probability of the other.

A critical feature is the fact that the highest amount of random information occurs when the outcomes, no matter how many possible, have equal probability of being received. This information is crucial, as the ambiguous option has the probabilities of the possible reward volumes hidden, the bits of random information are automatically assessed as at the maximum for the number of outcomes. Or in other words, the entropy of an ambiguous option where the number of possible outcomes is known, but the outcome probabilities are unknown, is treated as if the underlying distribution is uniform, and every outcome has an equal probability of being received, as this is the maximum amount of random information. This feature will uncover the specific effect of ambiguity, because if an ambiguous cue, and a risky cue (both with the same number of outcomes that are all equally probable) have the same amount of random information, reward range, and expected value, then any differences seen are only attributable to the specific effect of ambiguity.

Behavioral Task Cues to Measure Subjective Value of Uncertain Choice Options

We created a set of 4 risky cues and one ambiguous cue (Fig. 14a), which will all be offered as a choice versus a 'safe' certain cues that fully predict a reward. All of the cues use value bars, where a dot (or dots) intersect the vertical bar, indicating the reward volume(s) predicted, and the horizontal bar(s) stemming from the dot(s) indicate that reward volume's probability (Fig. 14a). The value range of the vertical line of the value bar represents is a range of 0-1 ml, spaced in 0.1 ml increments. The horizontal line, or summation of all the lines, will always total to 100%. Figure 14a shows the four risky, distribution predicting cues, and one ambiguous cue, with the top panel showing the low expected value (EV) conditions, where the reward distributions are drawn from 0.4, 0.5, and 0.6 ml of juice. The bottom panel of Figure 14a shows the high EV conditions, where the reward distributions are drawn from 0.7, 0.8, and 0.9 ml of juice. From left to right of Figure 14a, the Skew Low distribution (light blue) had a probability distribution of 0.50, 0.25, 0.25 for the low, medium, and high rewards. The Normal distribution (purple) had a probability distribution of 0.25, 0.50, 0.25 for the low, medium and high rewards. The Uniform distribution (green) had a probability distribution of 0.33 for all rewards. Because the maximum amount of uncertainty is when all options are equally probable (Fig. 12, Methods), the Ambiguous distribution (red) also had a probability distribution of 0.33 for all rewards, but with the horizontal probability bars occluded while the reward volumes were still visible. Finally, the Skew High distribution (dark blue) had a probability distribution of 0.25, 0.25, 0.5 for the low, medium, and high rewards. We used multiple reward distributions in order to create variety in the monkey's expectations of possible reward probability distributions that the occluder could be blocking. The Skew Low and Skew High distributions had slightly different EVs from the Normal, Uniform, and Ambiguous cues, and as such we solely use the Normal and Uniform cues as risky comparisons to the Ambiguous cue.

In each choice trial, after successful fixation of 0.5 s, the two choice options appeared on the monitor and the monkey indicated its choice by a saccade towards one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue of 0.5 s to confirm its choice. Reward was delivered 1.5 seconds later. In order to prevent the monkey from associating the grey occluder with a specific value, prevent a novelty response, and control for luminance across trials, it was always on the screen at variable positions, but only covered the reward probabilities of Ambiguous distribution trials. In the example trial in Figure 14b, the choice is between the low EV Normal distribution-predicting cue versus a safe value of 0.5ml. After selecting the Normal distribution-predicting cue, the reward would be drawn from the Normal distribution over 0.4, 0.5, and 0.6 ml of juice reward. The example trial in Figure 14c is between the high EV Ambiguous cue and a safe, certain value of 0.5 ml of juice reward. In this example trial, the reward would be drawn from a uniform distribution.

3.1.3 **Analysis of Behavioral Data**

Over 6 weeks, we collected ~5100 over 23 sessions choice trials of various uncertain distribution-predicting cues at two different EVs in one monkey. Due to the large variety of trial types (5 distribution cues, 2 different EVs, 11 ‘safe’ alternate options), we used week-by-week estimates of CEs to ensure enough trials to have a good fit of the psychometric functions. Due to the slight differences in EV in the Skew Low and Skew High cues, we removed them from further analyses – leaving the Ambiguous, Uniform, and Normal reward-predicting distributions for the low and high EVs.

3.1.3.1 **Calculating Subjective Value of Uncertain Options Using Certainty Equivalence**

Certainty equivalence (CE) is the value of a safe option that NHPs will have equal probability of choosing as an uncertain cue. In other words, the NHPs are indifferent between the two options Monkeys made choices between cues that predicted an uncertain cue and cues that explicitly indicated safe options (Fig. 14b, c). Figure 14d shows the probability of choosing the safe cue over the uncertain cue, as a function of the value of the safe option, when the distribution predicting cue had an EV of 0.5 ml (left), or 0.8 ml (right). Dots show average choice probability for 11 safe value options. Solid lines are a logistic fit to the data. The dashed horizontal lines indicate the point of certainty equivalence. Calculating the CE for all the uncertain options gave us a qualitative measure of how the animals valued the risky cues and ambiguous cue at low (0.5 ml) and high (0.8 ml) EVs. We used a Wilcoxon rank-sum test between Normal and Uniform distribution-predicting cues and found no significant differences for either EV (Fig. 15; Low EV, $p = 0.12$, High EV, $p = 0.62$; Wilcoxon rank-sum). This resulted reflects what we saw previously, that animals equally valued the Uniform and Normal distribution-predicting cues, and had matching CS responses in midbrain dopamine neuron firing rates, even though they had different underlying reward probabilities, in Chapter 2. We then compared the Uniform and Ambiguous cues, due to their matching EV, reward outcomes, and underlying distribution. There was a significant difference between Uniform and Ambiguous cues at both EVs, but the differences were in opposite directions. Specifically, at Low EVs, the monkey had a significant preference for Ambiguous cues over Uniform ones, based on their CEs ($p = 0.0001$, Wilcoxon rank-sum). However, in the High EV conditions, the monkey had a significant preference for Uniform cues over Ambiguous ones, based on their CEs ($p = 0.0043$, Wilcoxon rank-sum).

3.1.3.2 **Calculating Value-Dependent Uncertainty Preferences**

We plotted the CEs as a function of Low and High EV to determine the relationship between uncertainty-preferences as a function of expected value (Fig. 15b). When we compared the slopes of the Normal and Uniform risk-preferences from Low to High EVs, we see no significant difference, indicating their risk-preferences as a function of value were the same (Fig. 15b; Normal vs. Uniform slopes, $p = 0.75$; Wilcoxon rank-sum).

3.1.3.3 **Measuring Response Times**

Response time was considered the time between the cues being presented on the screen, and then selection of a cue. We compared response times between Uniform and Ambiguous cues with Low and High EVs. At the Low EV, the monkey showed equal response times during Uniform and Ambiguous choices (Fig. 15c $p = 0.75$, Wilcoxon rank-sum). At the High EV, the monkey took longer to make a decision in the Ambiguous choice conditions compared to Uniform ($p < 0.001$, Wilcoxon rank-sum).

3.1.3.4 **Deconvolution**

Event-related pupil responses were analyzed trial-by-trial using nideconv[236,237], a Python package that specializes in fMRI and pupil signal deconvolution. The design matrix for a trial consisted of a total of four event types: the onset of central dot for fixation, the onset of cue presentation, the monkey's saccades to indicate choice, offset of cue presentation (in temporal order), and the onset of reward. The pupil diameter changes related to fixation and the offset of cue presentation were analyzed 0.5 s pre-event until 2 s post-event; the time windows for the onset of cue presentation and monkeys' saccades started 0.5 s pre-event and ended 3 s post-event; the time window for the presence of rewards started at 0.5 s pre-event and ended at 1.5s post-event.

To understand the relationship between pupil diameter and uncertainty, expected value, reward volume, and reward prediction errors, were used as covariates in the deconvolution algorithm. Consequently, we obtained a measure of how sensitive the post-cue and post-reward pupil diameter changes were to the expected values, rewards received, and prediction errors in each trial, by looking at the beta coefficients in the prescribed time window.

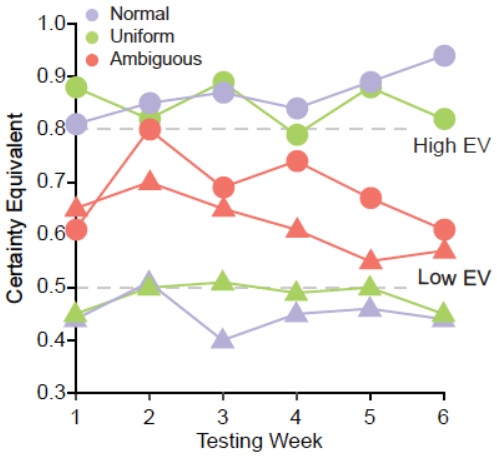


Figure 13: Certainty equivalent did not change over time
Certainty equivalent as a function of testing week, for Low EV versus High EV distribution-predicting cues. The Normal (purple), Uniform (green) and Ambiguous cues did not show any significant differences in CE over weeks of behavior and experience with the cues (Ambiguous, $p = 0.28$, Uniform, $p = 0.87$, Normal, $p = 0.94$; Kruskal-Wallis test). The circles represent CE at High EV, the triangles represent CE at Low EV. The bottom dashed line represents the EV of the Low EV distribution-predicting cues, and the top dashed line represents the EV of the High EV distribution-predicting cues.

3.1.3.5 Analysis of Possible Learning Over Multiple Weeks

In order to determine whether the monkey learned or updated the value of the Ambiguous cues over sessions, we performed a Kruskal-Wallis test to analyze the effect of week on each of the distribution-predicting cue types (Fig. 13). None of the cue types showed a significant effect of week on certainty equivalent, indicating that they did not learn over experience with the uncertain cues, and had stable CEs (Figure 13; Ambiguous, $p = 0.28$, Uniform, $p = 0.87$, Normal, $p = 0.94$; Kruskal-Wallis test).

3.2 **Results**

Decision-makers are ambiguity averse, and will choose cues with certain outcomes over ambiguous ones, at the expense of reward[1]. Previous studies have shown ambiguity aversion in non-human primates in choice situations with ambiguous versus risky cues[254]. However, it has not been determined how monkeys will evaluate the value of ambiguous options versus safe options, and whether the expected value of ambiguous options will change their tolerance of ambiguity, like what is seen with risk[269]. To investigate the specific effect of ambiguity as opposed to risk on decision making, we created a novel decision making paradigm with a variety of possible risky distributions over rewards, and the use of an occluder to block reward distribution information but not the possible values to create ambiguity.

We created a set of 4 risky cues with varying reward probability distributions and one ambiguous cue (Fig. 14a, Methods), which were all offered as a choice option versus a 'safe' certain cue that fully predict a reward. We used multiple reward distributions in order to create variety in the monkey's expectations of possible reward probability distributions that the occluder could be blocking. Further, to investigate the effect of expected value on ambiguity preferences, we utilized two different expected values for the uncertain cues, 0.5 ml EV (low EV) and 0.8 ml EV (high EV). In addition, the occluder was a constant feature of the task, even when it was not blocking any information on the value bars.

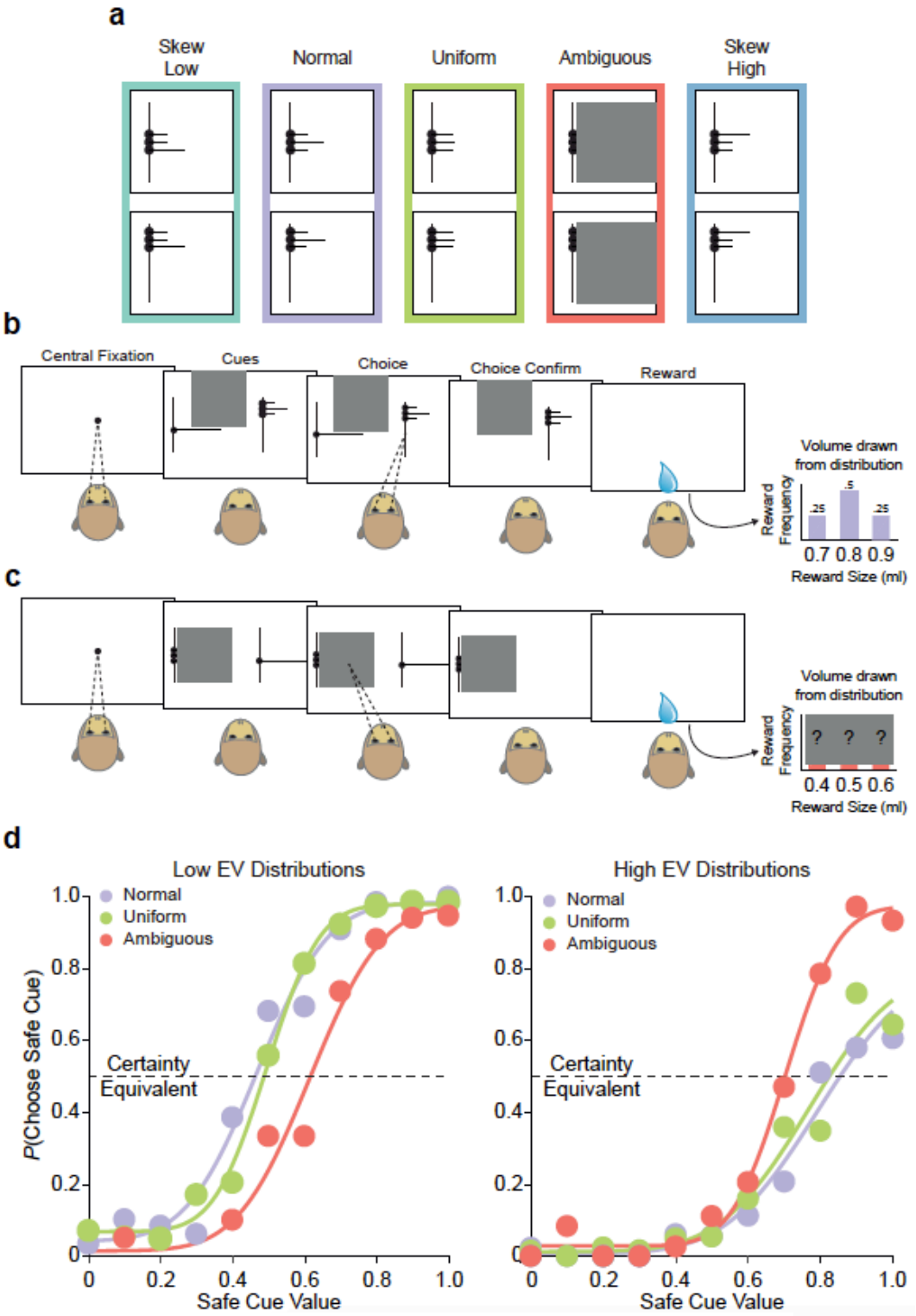


Figure 14: Choice behavior shows value-dependent risk and ambiguity preferences

a, Value bar cues to create four risky, distribution predicting cues, and one ambiguous cue in choice task to measure uncertainty preferences across two different expected values (EVs). The dots along the vertical line indicate the volume of reward, which are values between 0.1 and 0.9 ml in 0.1 ml increments. The horizontal bar that extends from the dot indicates the probability of that reward volume, with the sum of all the probabilities in a cue totaling to 1. The top row shows the low EV cues, with reward distributions over 0.4, 0.5, and 0.6 ml of juice. The bottom row shows

the high EV cues over 0.7, 0.8, and 0.9 ml of juice. Starting with the risky cues, the Skew Low (light blue) distribution cue reward probabilities were 0.50 for the low reward, and 0.25 for the medium and high reward. The Normal (purple) distribution predicting cue had probabilities of 0.25 for the low and high rewards and 0.50 probability for the medium reward. The Uniform (green) cue's rewards are equiprobable, and thus, have a probability of 0.33 for each. The Skew High (dark blue) distribution predicting cue reward probabilities were 0.25 for the low and medium rewards and 0.50 for the high reward. We used a variety of distributions in order to create variety in the monkey's expectations of possible reward probability distributions. In order to create ambiguity, we used a grey occluder the cover the reward probabilities while leaving the possible reward volumes visible. Because the maximum amount of uncertainty is when all options are equally probable (Fig. 12, Methods), the Ambiguous cue rewards were also drawn from a uniform distribution. The Skew Low and Skew High had slightly different EVs from the Normal, Uniform, and Ambiguous cues, and as such we solely use the Normal and Uniform cues as risky comparisons to the Ambiguous cue. **b,** Example trial of a medium EV Normal cue versus a safe, certain value of 0.5 ml of juice. In each choice trial, after successful fixation of 0.5 s, the two choice options appeared on the monitor and the monkey indicated its choice by a saccade towards one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue of 0.5 s to confirm its choice. Reward was delivered 1.5 seconds later. In this example trial, the reward would be drawn from the Normal distribution over 0.4, 0.5, and 0.6 ml of juice reward. In order to prevent the monkey from associating the grey occluder with a specific value, it was always on the screen at variable positions, but only covered the reward probabilities during the Ambiguous trials. **c,** Same as **b**, but for a high EV Ambiguous cue versus a safe, certain value of 0.5 ml of juice reward. In this example trial, the reward would be drawn from a uniform distribution (Methods). **d)** Probability of choosing the safe cue as a function of the value of the safe option, when the distribution predicting cue had an EV of 0.5 ml (left), or 0.8 ml (right). Dots show average choice probability for 11 safe value options. Solid lines are a logistic fit to the data. The dashed horizontal lines indicate certainty equivalence.

Because the maximum amount of uncertainty is when all options are equally probable (Fig. 12, Methods), the Ambiguous distribution (red) also had a probability distribution of 0.33 for all rewards, but with the horizontal probability bars occluded while the reward volumes were still visible. Finally, the Skew High distribution (dark blue) had a probability distribution of 0.25, 0.25, 0.5 for the low, medium, and high rewards. The Skew Low and Skew High distributions had slightly different EVs from the Normal, Uniform, and Ambiguous cues, and as such we solely use the Normal and Uniform cues as risky comparisons to the Ambiguous cue.

We first compared the Normal and Uniform distribution-predicting cues and found no significant differences for either EV (Fig. 15a; Low EV, $p = 0.12$, High EV, $p = 0.62$; Wilcoxon rank-sum). This resulted reflects what we saw previously, that animals equally valued the Uniform and Normal distribution-predicting cues in Chapter 2. We then compared the Uniform and Ambiguous cues, due to their matching EV, reward outcomes, and underlying distribution.

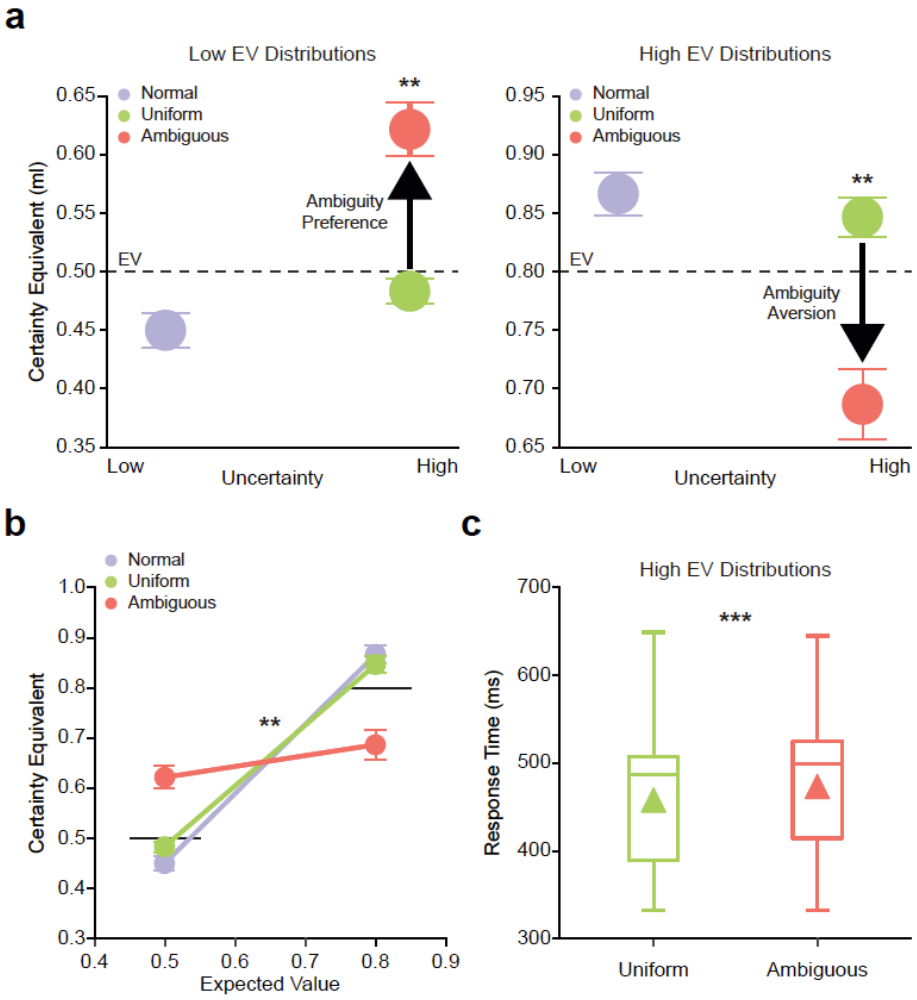


Figure 15: Ambiguity preferences are dependent on reward magnitude **a,** Certainty equivalent (CE) in ml of juice as a function of level of uncertainty for low (left) and high (right) EV distributions. This visualization shows that uncertainty was not a factor in the valuation of uncertain distribution-predicting cues, and that Ambiguity preferences were dependent on the expected value. The y axis represents the Certainty equivalent (CE) in ml of juice. The x axis represents the amount of uncertainty with Normal (purple) being lower than Uniform (green) and Ambiguous (red) cues. In low EVs, the Ambiguous cue had a higher CE than the Uniform cue (p = 0.0001, Wilcoxon rank-sum). This preference switched in the high EV conditions, and the Ambiguous cue had a lower CE than the Uniform one (p = 0.0043, Wilcoxon rank-sum). Normal and Uniform cue CEs were not significantly different in either EV condition (Low EV, p = 0.12, High EV, p = 0.62; Wilcoxon rank-sum). Error bars are ±s.e.m. across 6 week-by-week CE measurements. The horizontal dashed line represents the expected value (EV) of the cues. **b**, CE of uncertain cues across Low and High EV. Error bars are ±s.e.m. across 6 week-by-week CE measurements. Black horizontal lines indicate the expected value for the Low EV (0.5 ml) and High EV (0.8 ml) conditions. Slopes between Uniform and Ambiguous are significantly different ($p < 0.01$, Wilcoxon rank-sum). **c**, Box and whisker plots showing response times (ms) for Uniform (green) and Ambiguous (red) choices in the High EV condition. Ambiguous choices took significantly longer than Uniform ones ($p < 0.001$).

There was a significant difference between Uniform and Ambiguous cues at both EVs, but the differences were in opposite directions. Specifically, at Low EVs, the monkey had a significant preference for Ambiguous cues over Uniform ones, based on their CEs (Fig. 15a, left; p = 0.0001, Wilcoxon rank-sum). However, in the High EV conditions, the monkey had a significant preference for Uniform cues over Ambiguous ones, based on their CEs (Fig. 15a, right; p = 0.0043,

Wilcoxon rank-sum). This result shows that the Normal and Uniform cues had matched expected values but different distributions, it did not modulate their preference for either cue. However, even though the Uniform and Ambiguous distribution-predicting cues were matching in expected value and underlying distributions, the monkey showed variable preferences based solely on the fact that the Ambiguous cues had distribution information hidden from the animal. In addition, at the Low EV, the monkey showed equal response times during Uniform and Ambiguous choices ($p = 0.75$, Wilcoxon rank-sum). However, at the High EV, the monkey took longer to make a decision in the Ambiguous choice conditions compared to Uniform ($p < 0.001$, Wilcoxon rank-sum). We suggest that not only do they just completely avoid the ambiguous cues at High EVs, they deliberate longer.

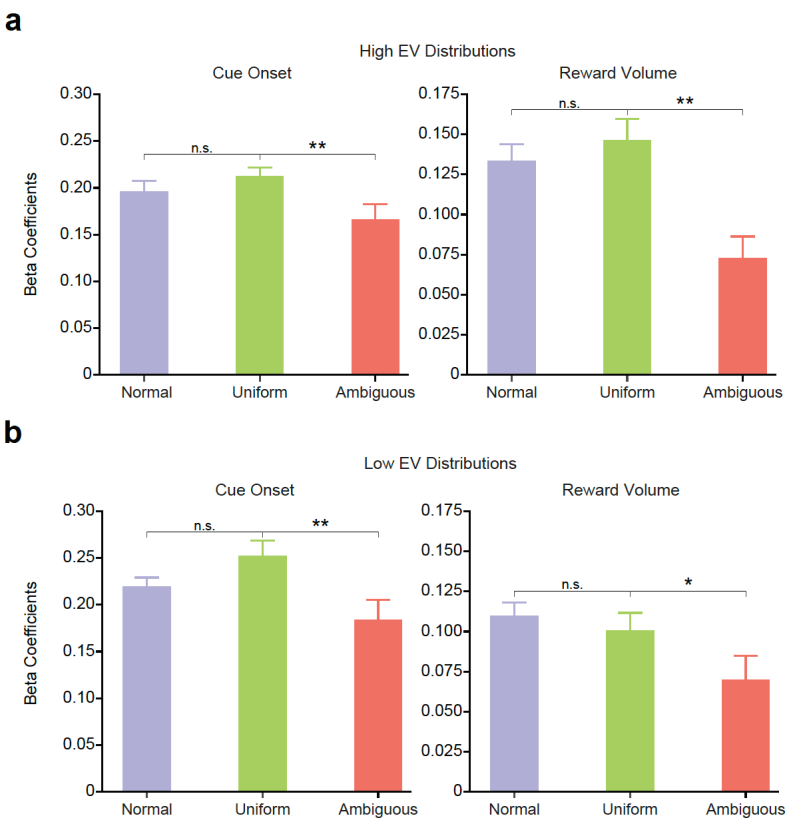


Figure 16: Ambiguous cues and thier following rewards illicit smaller pupil responses

a, Beta coefficients from a deconvolution analysis on the pupil diameter for Normal (purple), Uniform (green), and Ambiguous (red) cue onset (left) and reward reception (right) for High EV trials. **b**, Same as in **a**, but for Low EV trials. Error bars are ±s.e.m. across 23 sessions.

We used a deconvolution analysis to relate phasic changes of pupil diameter to different task epochs (Fig. 16). Across both EV conditions, pupil responses between the Normal and Uniform cues and their following rewards were never significantly different (Fig. 16a; High EV

Cue, $p = 0.22$, High EV Reward Volume, $p = 0.5$; Fig 16b; Low EV Cue, $p = 0.09$; Low EV Reward Volume, $p = 0.5$). In both High and Low EV conditions, we saw that phasic pupil responses were smaller following Ambiguous cues, and the rewards they received following them in compared to Uniform cues and their rewards (Fig. 16a; High EV Cue, $p < 0.01$, High EV Reward Volume, $p < 0.001$; Fig 16b; Low EV Cue, $p < 0.01$; Low EV Reward Volume, $p < 0.05$). Due to the reversal in preference for ambiguity from Low to High EV, this result shows that the pupil response was a value-independent response to ambiguity – such that the responses regardless of preference were always smaller relative to the Uniform cues and their subsequent rewards.

3.3 **Conclusion**

Decision-makers must assess the value of multiple options in order to make optimal choices. Most choices are made under some level of uncertainty, which are differentiated into two different types: risk and ambiguity[4]. Ambiguity aversion is the phenomenon where decision-makers will avoid ambiguity at the cost of utility[1,82,250-254]. Here, we show that NHPs show varying preferences for ambiguity which depends on the volume of the rewards – much like that of their preferences for risk[256]. Such that, when the stakes are low and rewards are small, they are ambiguity seeking, and when the stakes are high and rewards are small, preferences switch and the animals become ambiguity averse (Fig. 15a, b). While this study never directly gives animals choices between a risky and ambiguous choice option with the same EV, other research has shown ambiguity aversion specifically in relation to ambiguous versus risky options[254]. We hypothesize that at low stakes, animals will choose the ambiguous option more than a risky one, and at high stakes this will reverse. Future research offering animals these direct choices between risky and

ambiguous options in order to find an ambiguous cue's risky equivalent will be necessary to confirm this. In addition, at the High EV, the monkey took longer to make a decision in the Ambiguous choice conditions compared to Uniform (Fig. 15c). This result could indicate that not only are they ambiguity averse at the High EV, but they still take longer to decide, perhaps suggesting they take more time deliberating based on the uncertain or noisy estimates of EV in ambiguous situations.

We used a deconvolution analysis to relate phasic changes of pupil diameter to different task epochs (Fig. 16). We saw that phasic pupil responses were smaller following Ambiguous cues, and the rewards they received following them, in both High and Low EV conditions. Due to the reversal in preference for ambiguity from Low to High EV, this result shows that the pupil response was a value-independent response to ambiguity – such that the responses regardless of preference was the same relative to the Uniform cues. This reveal a specific relationship with pupil diameter and ambiguous economic choices. Some research has shown that phasic changes in pupil diameter changes have been linked to unexpected uncertainty, which is also similar to novelty detection and the necessity to update the underlying probability of the environment[266,270-272]. However, we believe the uncertainty from ambiguity experienced in our task is not reflective of unexpected uncertainty, as the occluder is not novel in any way, and the underlying task statistics of the environment are held constant. Other research has shown that faster learning rate is reflective of an increased state of arousal, reflected in phasic increases in pupil diameter[273]. In Chapter 2, we found evidence supporting this, specifically that less uncertainty in the distribution of reward outcomes associated with a cue elicited faster learning and increases of phasic pupil responses. Here, there was no learning in relation to any of the cues (Fig. 13), and the Ambiguity cue and its

rewards elicited smaller phasic pupil responses, which is suggestive that the ambiguity elicited by the occluder was not perceived as novel by the animal.

The monkey could potentially attempt to learn, through trial-and-error[9], the underlying reward distribution of the ambiguous cue. It is the same in reward outcome possibilities and probabilities, uncertainty, and EV as the Uniform distribution-predicting cue, with the only difference being that the probability information is hidden from the animal. Indeed, previous work with ambiguity has shown that monkeys did eventually learn the underlying probabilities of rewards through multiple experiences with an ambiguous cue[254]. Further, in Chapter 2, we show that animals can indeed learn the expected values of non-informative (fractal) Uniform distribution-predicting cues, albeit slower than fractal Normal distribution-predicting cues[267]. While we did not have the same variety of underlying distributions in Chapter 2 that we did here, the overall frequency of all the rewards, if you consider the probabilities of all the risky cues, is equal across the three reward volumes of the distributions. Or in other words, if the animal were to have integrated information about the overall distribution of the environment to estimate the most likely underlying distribution, it could have inferred a uniform distribution of rewards as well. Considering these possible methods that the animal could have used to learn the underlying distribution, either through trial-and-error experience or by integrating information about the overall uncertainty of the environment, leads us to believe that the specific differences seen here between the Uniform and Ambiguous cues can only be attributed to the specific effect of ambiguity, seen in humans and animals[1,82,250-254]. Together, our results suggest that different psychological and neural mechanisms mediate risk and ambiguity preferences in reward-based economic decision making. This view is consistent with other studies that have suggested the same[165,254,274].

It has been shown that learning about the Ellsberg Paradox reduces, but does not totally abolish, ambiguity aversion, and this increases behavioral performance in humans[255]. This finding shows that while this is somehow an engrained bias individuals have; cognitive inhibition of this behavior can be implemented. This leaves room to investigate what brain areas perform top-down modulations to economic decision making in choices with ambiguity. Some areas that have previously been shown to be differentially responsive to ambiguity versus risk in economic decision making are the lateral orbitofrontal cortex (lOFC) and the amygdala. A previous study using fMRI in humans performing a decision making task show increases in lOFC and the amygdala in ambiguous decisions compared to risky ones[274]. Further, the striatum was more activated in risky decisions over ambiguous ones. These results provide the basis for further behavioral and neuronal characterizations of decision making under ambiguity.

4.0 Transcriptional and Anatomical Diversity of the Primate Striatum[2]

The striatum serves as the major input nucleus for the basal ganglia (BG) and the principal neural interface between dopamine reward signals and cortico-basal ganglia-thalamo-cortical circuits. Information processing in the striatum is dependent on cell-type-specific circuits. In particular, Medium Spiny Neurons (MSNs), which account for the vast majority of all striatal neurons, are divided into two major cell types: D1- and D2- MSNs[275]. D1-MSNs express dopamine receptor type 1 (*DRD1*) and form the "direct pathway" via monosynaptic projections to the basal ganglia output nuclei[103]. D2-MSNs express dopamine receptor type 2 (*DRD2*) and form the "indirect pathway" via di-synaptic projections to the basal ganglia output nuclei[103]. Activity in the direct and indirect pathways produces, broadly, opposing effects on thalamo-cortical projections. This cell-type-specific circuit model has been crucial to understanding the role of the striatum and BG in the control of movement and the mechanisms of Parkinson's disease[276,277]. However, the striatum and BG are involved in many behaviors besides the control of movement. For example, segregated neurochemical compartments in the dorsal striatum (DS), known as striosome (patch) and matrix are thought to participate in limbic and sensorimotor functions, respectively[278-280]. Similarly, the ventral striatum (VS) has fundamental roles in reward processing, learning, and emotional responses[281-284]. The traditional model, that involves competition between signals in the direct and indirect pathways, does not account for these broad functional roles. Rather, these broad functionalities indicate deeper cell-type and circuit heterogeneity.

[2] *The contents of this chapter were previously published (He & Kleyman et al., 2021). KMR made significant intellectual contributions in the planning and execution of data acquisition for the manuscript.*

Single cell technologies that classify cell types according to their overall gene expression profiles provide powerful and quantitative methods for investigating cell type heterogeneity[285]. Single cell and single nucleus RNA sequencing (scRNA-Seq and snRNA-Seq, respectively) have revealed new subtypes of MSNs and striatal interneurons[286-291]. Moreover, these technologies are providing novel insights into cell-type-specific mechanisms for diseases involving the striatum, including drug addiction[292] and Huntington's disease[293]. Despite these advances, we know neither the extent of MSN diversity in the primate striatum, nor how that diversity corresponds to the anatomical or neurochemical divisions of the highly-articulated primate brain.

The close phylogenetic relationship and the high degree of homology between human and non-human primate (NHP) brains, genes, and behaviors make NHP studies indispensable for understanding the neuronal substrates of human behavior, as well as neurological, neurodegenerative, and psychiatric diseases[294]. Here, we used snRNA-Seq and Fluorescent In-Situ Hybridization (FISH) to characterize the transcriptional and anatomical diversity of MSNs and closely related neurons. The resulting cell-type-specific gene expression patterns provide insights into MSN functions and indicate potential molecular access points for cell-type-specific applications of genetically coded tools to primate brains, in scientific or translational settings.

4.1 **Methods**

4.1.1 **Non-human Primates (NHPs)**

All animal procedures were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the University of Pittsburgh's

Institutional Animal Care and Use Committee (IACUC) (Protocol ID, 19024431). Rhesus Monkeys were single- or pair-housed with a 12h-12h light-dark cycle. Monkey F was a 12-year-old female (8.1 kg). Monkey P was a 5-year-old female (5.4 kg). Monkey B was a 13-year-old male (11.78 kg). Monkey K was a 4-year-old male (6.0 kg).

4.1.2 **MRI and Surgery**

For MRI, we anesthetized monkey F and P with ketamine and maintained general anesthesia with isoflurane. We head fixed the monkeys using an MRI-compatible stereotaxic frame and scanned (Siemens, 3T) for anatomical MRI. We generated a whole brain model for each monkey using Brainsight (Rogue Research) and 3-D printed a custom matrix for the brain with cutting guide set every 1 mm.

To maximize nuclei viability, we followed a harvesting protocol similar to the one outlined in Davenport et al.[295]. Briefly, animals were initially sedated with ketamine (15 mg/kg IM), and then ventilated and further anesthetized with isoflurane. The animals were transported to a surgery suite and placed in a stereotaxic frame (Kopf Instruments). We removed the calvarium and then perfused the circulatory systems with 3-4 liters of ice cold artificial cerebrospinal fluid (ACSF; 124 mM NaCl, 5 mM KCl, 2 mM $MgSO_4$, 2 mM $CaCl_2$, 23 mM $NaHCO_3$, 3 mM NaH_2PO_4, 10 mM glucose; pH 7.4, osmolarity 290–300 mOsm) oxygenated with 95% O_2:5% CO_2. We then opened the dura and removed the brain. We sliced on the custom brain matrix into 4 mm slabs along the rostral-caudal axis. We removed three striatal regions – the caudate nucleus, putamen, and ventral striatum –under a dissection microscope for nuclei isolation. Monkeys B and K, for FISH, were perfused with 4% paraformaldehyde (PFA, Sigma-Aldrich, Cat# P6148) in phosphate buffered saline (PBS, Fisher Scientific, Cat# BP243820) supplemented with 10% sucrose (Sigma-

Aldrich, Cat# S8501). The brain was post-fixed with 4% PFA and cryopreserved with a gradient of sucrose (10%, 20%, 30%) in PBS.

4.1.3 **Nuclei Isolation**

We isolated nuclei isolated as previously described[296]. Briefly, we homogenized tissues using a loose glass dounce homogenizer followed by a tight glass homogenizer in EZ PREP buffer (Millipore Sigma, Cat# NUC-101). We washed nuclei once with EZ PREP buffer and once with Nuclei Suspension Buffer (NSB; consisting of 13PBS, 0.01% BSA and 0.1% RNase inhibitor (Clontech, Cat# 2313A)). We re-suspended the washed nuclei in NSB and filtered them through a 35-mm cell strainer (Corning, Cat# 352235). We counted the nuclei and diluted down to 1000 cells/ml. We loaded approximately 10,000 cells from each brain region onto a 10X chip which were then run through a 10x Genomics Chromium controller.

4.1.4 **Single Nucleus RNA-Seq**

We used 10x Chromium Single Cell 30 Reagent kits v3 Chemistry (10x Genomics, Cat# PN-1000075) for monkeys F and P. We reverse transcribed RNAs and generated libraries according to 10x Genomics protocol. Briefly, we generated Gel beads-in-emulsion (GEMs) after running through a 10x Genomics Chromium controller. We reverse transcribed mRNAs within GEMs in a Bio-Rad PCR machine (Cat# C1000). We barcoded cDNAs from individual cells with 10x Genomics Barcodes and barcoded different transcripts with unique molecular identifiers (UMIs). We purified cDNAs with Dynabeads (10x Genomics, Cat# 2000048) after breaking the emulsion with a recovery agent (10x Genomics, 220016). Then, we amplified cDNAs by PCR and

purified them with SPRIselect reagent (Beckman Coulter, Cat# B23318). We analyzed the cDNA quantification and quality using Agilent Bioanalyzer 2100. We prepared libraries following fragmentation, end repair, A-tailing, adaptor ligation, and sample index PCR. We quantified the libraries by qPCR using a KAPA Library Quantification Kit (KAPA Biosystems, Cat# KK4824). We pooled together libraries from individual monkeys and loaded them onto NovaSeq S4 Flow Cell Chip. We sequenced samples from monkeys F and P to depths of 400,000 and 250,000 reads per nuclei, respectively.

4.1.5 **FISH Probes**

We ordered custom FISH probes from ACD to validate MSN subtypes as follows: *DRD1* (ACD #549041, a 20ZZ probe targeting 1335-2279 of NM_001206975.1), *DRD2* (ACD #549031-C2, a 20ZZ probe targeting 232-1470 of XM_001085571.3), *RXFP1* (ACD #801121-C2, a 20ZZ probe targeting 1508-2592 of XM_001096574.4), *CPNE4* (ACD #801111-C3, a 20ZZ probe targeting of 2-943 of XM_028843706.1), *KCNIP1* (ACD #889143-C3, a 20ZZ probe targeting 283-1626 of XM_015141392.2), *KCNT1* (ACD #881571-C3, a 20ZZ probe targeting 964-1906 of XM_015116381.2), *BACH2* (ACD #898961-C3, a 20ZZ probe targeting 927-2260 of XM_028847343.1), *KHDRBS3* (ACD #881591-C3, a 20ZZ probe targeting 494-1493 of XM_028852934.1), *STXBP6* (ACD #881611-C2, a 20ZZ probe targeting 2-1077 of NM_001260925.1), *SEMA3E* (ACD #879971-C2, a 20ZZ probe targeting 905-1889 of XM_028846126.1), *GDA* (ACD #881601-C2, a 20ZZ probe targeting 355-1373 of XM_015117899.2), *GREB1L* (ACD #898991-C3, a 20ZZ probe targeting 784-1732 of XM_015121640.2), *ARHGAP6* (ACD #898981-C3, a 20ZZ probe targeting 2516-3463 of XM_001094565.3), *TAC3* (ACD #520901-C3, a 17 ZZ design and targets 2- 768 bp of

XM_001115535.2), and *OPRK1* (ACD #518931, a 20 ZZ probe and targets 91-1392 bp of NM_001321097.1).

4.1.6 **FISH Stain and Imaging**

We embedded brains in optimal cutting temperature (OCT) and stored them at 80°C until cutting. We cut floating sections at 15 and 30 mm, in monkey B and K, respectively, mounted tissue on 2x3" pre-treated slides, and preserved the slides in a freezer at -80°C. We used the Advanced Cell Diagnostics (ACD) RNAscope platform and Multiplex Fluorescent Detection Reagents v2 (ACD, Cat# 323110) to perform FISH with slight modifications for monkey brain tissue. We air-dried slides for 30 min after removal from -80°C freezer and baked them at 60°C for 20 min. We treated the brain sections with hydrogen peroxide (ACD, Cat# 322335) for 10 min at room temperature and then with RNAscope Target Retrieval Reagents (ACD, Cat# 322000) for 8 min at 99°C. We incubated the slides in 100% alcohol for 3 min and then dried them in 60°C for 10 min. We treated the samples with protease III (ACD, Cat# 322337) for 30 min at 40°C and incubated them with probes for 2 hr. After hybridization with AMP 1, 2 and 3, we incubated the slides with different HRP channels and fluorophores, including Opal 520 (PerkinElmer, Cat# FP1487A), Opal 570 (PerkinElmer, Cat# FP1488A) and Opal 650 (PerkinElmer, Cat# FP1496A). Lastly, we used Trueblack (Biotium, Cat# 23007) to quench Lipofuscin autofluorescence for 45 s at room temperature and counterstained every slide with DAPI before mounting with Prolong Gold Antifade Mountant (Life technologies, Cat# P36930).

We scanned labeled sections using a Hamamatsu NanoZoomer S360 or a Nikon Eclipse T*i*2 under 20x objective. For high resolution images, we used a Nikon Eclipse Ti2 or an Olympus

IX83 under 40x or 63x objective or 40x with additional 2x built in objective (equal to 80x). We took multi-layer images and did deconvolution using Nikon's NIS-Elements deconvolution software and used maximum intensity projections to create single images. We used NDP software to convert the NanoZoomer original files to tiff format and ImageJ and Adobe Photoshop to adjust brightness and overlay images.

4.1.7 **Immunohistochemistry (Fluorescent)**

We sought to verify previously labeled sections containing FISH probes for possible shell markers using adjacent tissue sections and immunohistochemistry to label for Calbindin. The sections were rinsed in Phosphate Tris (PT, pH 7.2-7.4) buffer and blocked in 10% normal donkey serum (Jackson, Cat# 017-000-121) solution for one hour. The sections were then incubated with primary antibody solution (Swant, Calbindin D-28K, 300, 1: 7,000) overnight at 4°C. The sections were then rinsed in PT buffer and incubated in secondary solution (Alexa Fluor 568, Donkey anti-mouse, 1:300) at room temperature for two hours. They were then rinsed in PT buffer and counterstained with Hoechst (Invitrogen, Cat# H21486, 1: 10,000) for 10 minutes. Lastly, the sections were rinsed in PT buffer and mounted with Prolong Gold Antifade (ThermoFisher, Cat# P36930). To verify μ-opioid receptor proteins were enriched in NUDAPs as well, we did immunostaining with MOR antibodies on an adjacent section performed with *OPRM1* FISH labeling (Abcam, ab-10275, 1:500) using similar approach except that PBS buffer instead of PT buffer was used. To verify *DRD2* and *CPNE4* labels cholinergic neurons, after FISH labeling with the two probes, we performed the following immunostaining with ChAT antibodies (Pro-Sci, Cat# 45-037, 1:1000) using similar approach except that TBS buffer instead of PT buffer was used.

4.1.8 **Immunohistochemistry (DAB)**

Free-floating sections were selected based on previous FISH labeling to provide verification of common protein marker identification. Sections selected for KChIP1 were rinsed in Phosphate Tris (PT, pH 7.2-7.4) buffer and Phosphate-Buffered Saline with 0.2% Triton (PBST, pH 7.2-7.4) respectively. They were then moved into a 0.5% H_2O_2 buffer for 10 minutes followed by a series of rinses before incubating in a 10% Normal Horse Serum (NHS, Vector, S-2000) solution for one hour. The tissue was transferred to primary antibody solution (NeuroMab, Anti-KChIP1, clone K55/7, 1:200) for overnight incubation at 4°C. The next day, sections were again serially rinsed in respective buffers and incubated in secondary antibody (Vector, Vectastain ABC, Peroxidase Kit, PK-4002, 1:200) for 30 minutes. Sections were again serially rinsed and placed into ABC (Vector, Vectastain ABC, Peroxidase Kit, PK-4002) buffer for one hour incubation before being rinsed. Tissue was placed in DAB substrate solution (3,3′-diaminobenzidine, Vector, SK-4100) until reacted. Following a final serial rinse, sections were mounted for imaging and analysis.

See also Data S1, S2, S3, and S4.

4.1.9 **Quantification and Statistical Analysis**

4.1.9.1 **Custom Annotation File**

We downloaded the macaque rheMac10 genome[297] (https://hgdownload.soe.ucsc.edu/goldenPath/rheMac10/bigZips/rheMac10.fa.gz) and human NCBI RefSeq transcriptome annotation gtf file from the UCSC genome browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/hg38.ncbiRefSeq.gtf.gz). We

used the UCSC liftOver tool with the hg38toRheMac10 chain file (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/liftOver/hg38ToRheMac10.over.chain.gz) to overlay the human transcriptome gtf file onto the rheMac10 genome, with a minimal match threshold of 0.85. The human "liftOver" annotations were used to extend and supplement the macaque annotations, leading to greater numbers of genes called (Figure S1B).

4.1.9.2 **Single Nucleus RNA Sequencing Analysis**

We converted original bcl format of sequences to fastq files using cellranger mkfastq. Alignment of the reads by cellranger count to the rheMac10 genome using the custom rheMac10 gtf yielded 61,609 and 23,690 nuclei for monkey P and F, respectively. We used a Seurat v3 pipeline to perform an integrated analysis of monkey F and P (85,299 nuclei in total). First, we removed ambient RNA in monkey P using SoupX[298]. We then deleted ribosomal genes for both monkeys, and removed doublets using DoubletDetection[299], which lowered the number of nuclei to 80,902. Neuronal cells express higher numbers of genes compared to non-neuronal cells[300-302] and therefore we used two different thresholds to remove low quality neuronal and non-neuronal nuclei. We chose these thresholds as the minimal level that produced clear cluster separation (Figure 17E). A total of 31,258 genes were identified, with an average of 3,200 per nucleus. We performed standard log-normalization and a variance stabilizing transformation prior to finding anchors and identified variable features individually for each monkeys' dataset using Seurat's *FindVariableFeatures* function with number of features set to 7000. Next, we identified anchors using the *FindIntegrationAnchors* function with default parameters and passed these anchors to the *IntegrateData* function. This returned a Seurat object with an integrated expression matrix for all nuclei. We scaled the integrated data with the *ScaleData* function, ran PCA using the *RunPCA* function, and visualized the results with UMAP. We used Louvain clustering and chose a

resolution that reflected the major cell classes of striatum including D1- and D2-MSNs, interneurons and astrocytes. We calculated the differentially expressed genes for each cell class with the *FindMarkers* function (Data S1). We annotated the clusters based on feature plots of well-known marker genes[286,290,303] and verified the identities of cell clusters by using the hypergeometric test to compare differentially expressed genes for each cluster to markers from single cell rodent studies[304]. Given the robust conservation of major cell types, the rodent markers were sufficient for annotation[304]. We converted rodent genes to rhesus macaque genes by BioMart Ensemble, keeping one-to-one orthologs only[305,306].

For MSN analysis, we isolated the clusters that were enriched with well-known marker MSN genes including *PPP1R1B*, *BCL11B*, and *PDE1B*. In order to balance the number of nuclei per animal, we randomly sampled MSN and MSN-like nuclei from monkey P to levels of monkey F (7,387 nuclei). We then re-calculated principal components (PCs) and performed UMAP dimensionality reduction on the first 15 PCs. We used Louvain clustering and chose a resolution that separated clusters that were distinct in UMAP space. We calculated the differentially expressed genes for each MSN subtype with the *FindMarkers* function (Data S1). To determine whether the clustering of MSNs was exhaustive, we further isolated 'D1-' and 'D2-MSNs'. D1-MSNs including D1-striosome, D1-matrix and D1-shell/OT, whereas D2-MSNs included D2-striosome, D2-matrix and D2-shell/OT. We re-calculated PCs and performed UMAP dimensionality reduction based on the top 15 PCs for both subclusters. This analysis recovered the same, physically distinct clusters observed in the integrated analysis. To analyze the interneuron populations, we isolated clusters based on interneuron markers[291]. Similarly, we re-calculated PCs and performed UMAP dimensionality reduction on the first 30 PCs. We used Louvain clustering and annotated the resulting clusters based on known interneuron markers. To explore the

functional roles of these cell types, gene enrichment analysis was run using gprofiler[307] with the following GO databases: Biological Process, Molecular Pathway, KEGG, and Human Phenotype ontology (Data S4). We calculated the cosine similarity for nuclei within and between clusters based on PCA space. We used a permutation test on the cosine similarity between pairs of clusters. We randomly shuffled the nuclei to mask the nuclei identity and recalculated cosine similarity. We repeated these 10,001 times and used within group and between group variance ratio to determine a p value. To compare the cell types between monkey and mouse MSNs, we first generated an orthologous gene list between mouse and rhesus macaque from BioMart Ensemble with one-to-one orthologous genes[305,306] used for downstream analysis. We integrated our MSNs and MSN-like nuclei from the two monkeys with MSN types in Stanley et al.[288] based on the homologous genes using the Seurat integration method as above for the two animals.

4.1.9.3 **Archetypal Analysis**

We used archetypal analysis to characterize the gradients within and between subtypes. We used partition-based graph abstraction (PAGA) to find those pairs that warranted a gradient-based analysis. PAGA uses the formalism of graph theory and community detection to define a statistic quantifying the presence of connectivity between two clusters[308]. To calculate the PAGA graph we ran *scanpy.tl.paga* on our Seurat integrated data (Figure S3F). We defined subtype pairs as connected if the PAGA edge weight was greater than 0.02. For each pair of connected subtypes – or within a single subtype – we used the Dirichlet Simplex Nest (DSN) implementation of archetype analysis (https://arxiv.org/abs/1905.11009) to define gradients of gene expression. Because we used the raw gene counts, we ran the DSN algorithm in the Poisson configuration[309]. Across several runs of the DSN algorithm, we located the archetype most correlated with the subtypes' annotations and designated those archetypes as the transition axes between the subtype

pairs. To determine if the transitions were discrete, we used regression discontinuity design (RDD)[310]. The null hypothesis of this test is that the gene expression is, solely, a linear function of the transition archetype weights, and adding a threshold does not add additional information. Therefore, we computed p value distributions across all correlated genes for each connected pair. We then tested whether the number of genes that had significant discontinuities ($p < 0.05$, RDD) was greater than expected by chance using a binomial test. We performed the binomial test with two different null distributions, a uniform null distribution, and a simulated null distribution where we randomly shuffled the cell labels and performed the RDD 100 times per gene. Regardless of which null distribution we used, the binomial tests indicated more statistically significant discontinuities than predicted by chance for every subtype pair except for D1S and D2S (the subtypes for which we had the lowest sample sizes) ($p < 0.00001$ for uniform null, and $p < 10^{-12}$ for simulated null). We corrected the binomial test p values using the Benjamini-Hochberg procedure. To verify biological reproducibility, and because defining an archetype and computing discontinuities from the same data can lead to inflated p value distributions[311], we calculated the archetypes using data from monkey P and computed the discontinuities on data from monkey F. To project the archetypes learned from monkey P onto monkey F, we multiplied the normalized expression matrices of monkey F by the pseudoinverse of the archetype vectors. Using the DSN algorithm we found several archetypes of particular interest including an archetype representing *CNR1* signal in hybrid cells, caudate signal in D1- and D2- MSNs, core signal in D1- and D2-MSNs and *TAC3* signal in shell D1- MSNs. To provide molecular markers for these archetypes we calculated each gene's Pearson correlation with each archetype of interest (Data S2 and S3; Tables S2 and S3).

4.1.9.4 **Assessing Clustering Robustness**

We used Single Cell Clustering Assessment Framework (SCCAF)[312] to test the robustness of our MSN subtype classification. The concept behind such 'self-projection' tests is that the gene expression patterns in a subsample of the cells should be sufficient to classify the remaining cells in the labeled clusters with a high level of accuracy[313]. SCCAF splits the expression data into a training and test set, and then fits a classifier on the training set on each cluster provided (MSN subtypes in our case). We used the default parameters of a 50% train/test split and a logistic regression classifier. We also used SCCAF to test whether our snRNA-seq sampling was sufficient for accurately detecting the MSN subtype heterogeneity. Because the number of cells of each MSN subtype could vary during the down sampling, we opted to use SCCAF classifier's macro-f1 score for evaluation. The f1 score is the geometric mean between the precision and recall. To compute our macro-f1 score, we computed the average across f1 scores for each MSN subtype. Using the *scanpy.pp.subsample* method, we repeatedly subsampled fractions of our cells from both monkeys and calculated the macro-f1 across each trial.

4.1.9.5 **FISH Image Quantification**

We quantified expression using high resolution images collected using a Nikon Eclipse Ti2 (40x objective with or without additional 2x built in objective) or an Olympus IX83 under 40x or 63x magnification. For comparison between different regions, we scanned images using the same settings. To quantify cells expressing *DRD1* and *DRD2* in the caudate and putamen (Figure 20D), we chose ten representative areas from each striatal region (Figure 20C). We adjusted the threshold of the nuclei images and converted them to black and white using the "Make Binary" function in ImageJ[314]. We then filled the holes using the "Fill Holes" function and separated overlapping nuclei with the "Watershed" function. We identified the number and regions for nuclei using the

"Analyze Particles" function and added the nuclear contours to "ROI Manager." We then opened the *DRD1* and *DRD2* images and identified the *DRD1* and *DRD2* grains using the "Find Maxima" function in ImageJ and then obtained binary images with a single pixel for each local maxima by choosing output type of "Single Points." We measured the integrated intensity of *DRD1* and *DRD2* above each nucleus using the "Measure" function within the "ROI Manager." The number of grains in each nucleus is equal to the integrated intensity divided by 255. We considered cells containing three or more grains above the nucleus as positive for that gene. We chose this threshold because it provides clear separation from the background. We quantified the total number of *DRD1*-positive and *DRD2*-positive cells for each striatal region of interest (ROI) in each section and counted positive cells on three rostro-caudal sections. We calculated cell density as the number of positive cells divided by total number of nuclei in the area. We used a similar method to quantify the cell density in *CPNE4* and *RXFP1* clusters (Figure S6E) except that we drew ROIs for the whole *CPNE4* and *RXFP1* clusters after separating overlapping nuclei with the "Watershed function. As control, we drew ROIs of roughly similar size to the *RXFP1* or *CPNE4* clusters in nearby regions and quantified the cells expressing *DRD1*.

To quantify *DRD1* and *DRD2* grains in D1/D2-hybrid cells, D1- and D2- MSNs (Figure 20G), we scanned high resolution images for *RXFP1*-positive cells in triple stained *DRD1*, *DRD2*, and *RXFP1* sections. The majority of *RXFP1*-positive cells expressed both *DRD1* and *DRD2* in dorsal striatum. We quantified the number of grains for *DRD1* and *DRD2* in these D1/D2-hybrid cells as well as adjacent normal D1 and D2 MSNs in ImageJ using similar methods as above except that we quantified the total grains in the cells instead of nuclei. To quantify *DRD1* and *DRD2* grains in D1/D2-hybrid cells, we first draw ROIs for *RXFP1* expressing cells and added the ROIs to the "ROI Manager" based on the *RXFP1* signal. We then opened the *DRD1* and *DRD2* images

and identified the *DRD1* and *DRD2* grains using the "Find Maxima" function in ImageJ and then output binary images with a single pixel for each local maxima by choosing an output type of "Single points." We measured the integrated intensity and calculated the number of grains for each ROI using the "Measure" function within the "ROI Manager." To quantify *DRD1* and *DRD2* grains in D1- and D2- MSNs, we used the same method except that we drew ROIs for D1- and D2- MSNs. We confirmed the identity of D1- or D2- MSN instead of a D1/D2-hybrid cell by quantifying the *DRD2* or *DRD1* and *RXFP1* grain number less than three in D1- or D2- MSNs. We used a similar method to quantify *ARHGAP6* and *GREB1L* grains in the core and shell except that we first adjusted the threshold of the images to overexpose the *ARHGAP6* and *GREB1L* signals to guide the ROI drawing. To quantify *OPRM1* grain numbers in *RXFP1* and *CPNE4* islands and nearby D1-MSNs, we labeled *DRD1* and *OPRM1* on one section and *DRD1*, *RXFP1* and *CPNE4* on an adjacent section because of overlapping channels in the *OPRM1* and *RXFP1* probes. The locations of *RXFP1* and *CPNE4* islands in the *OPRM1* and *DRD1* double stained sections were determined by adjacent section labeled with *RXFP1* and *CPNE4*. The quantification of *OPRM1* grains numbers was similar to *ARHGAP6* or *GREB1L* quantification except that we used overexposed *DRD1* to draw ROIs for each cell.

To quantify nuclei size for *RXFP1* and *CPNE4* islands (Figure S6F), we first scanned high resolution images for *DRD1*, *RXFP1*, and *CPNE4* triple labeled sections. We performed automatic quantification of the areas of nuclei using ImageJ. We adjusted the threshold and convert images to black and white using the "Make Binary" function. Then filled the holes using the "Fill Holes" function. Then we separated overlapping nuclei with the "Watershed" function, and then draw the ROI and automatically detect the areas of nuclei using the "Analyze Particles" function. To produce a more accurate quantification of the nuclei size, we chose to randomly sample dozens of

cells expressing *DRD1* in each island per section due to high packing density of the nuclei in ICjs (Figure S6E). We calculated the area of nuclei by the similar method except that we drew an area of interest around individual nucleus. Three sections in total were quantified and the area for each nucleus was normalized to the control D1-MSNs in each section.

4.1.9.6 **D1 Islands Mapping**

In order to map the distribution of D1-*RXFP1* and D1-*CPNE4* islands, we triple labeled six sections spaced at 750 mm intervals with probes against *DRD1*, *RXFP1*, and *CPNE4* in monkey B. We scanned the whole sections and ran a custom CellProfiler pipeline[315] to determine the spatial distribution of nuclei and signal intensity of individual channel for each nucleus. We marked regions as D1-*RXFP1* (D1-*CPNE4*) that had a majority of cells double-labeled with *DRD1* and *RXFP1* (*CPNE4*). To confirm that these islands were D1 exclusive, we double-labeled sections with *DRD1* and *DRD2* adjacent to the above six sections. We confirmed that these islands are exclusively the D1 clusters from the similar CellProfiler analysis. We did the similar process for eight sections spaced at 600 mm intervals for monkey K.

4.2 **Results**

4.2.1 **Major Cell Classes in the Primate Striatum**

To investigate the cell-type-specific architecture of the NHP striatum, we micro-dissected the caudate nucleus (Cd), putamen (Pt), and ventral striatum (VS) from coronal sections of two monkeys (Figures 17A and S1A), performed snRNA-Seq, and clustered the nuclei profiles based

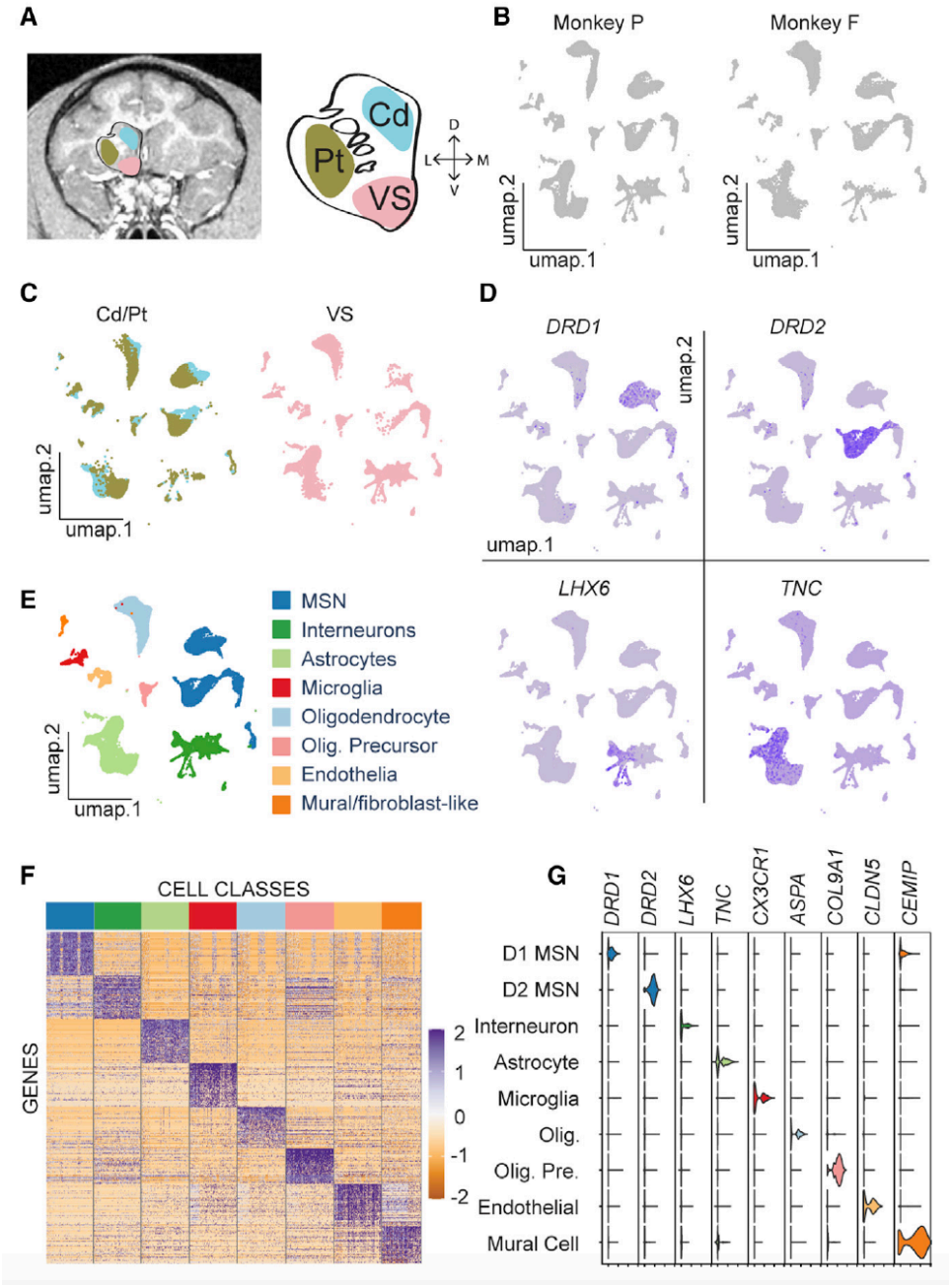


Figure 17: Cell type taxonomy in the primate striatum

A, MRI image of a Rhesus macaque coronal brain section (left) showing three striatal regions labeled with cyan (caudate nucleus [Cd]), brown (putamen [Pt]), and pink (ventral striatum [VS]). Schematic striatum (middle) marked by Cd, Pt, and VS. The right axis shows dorsal (D), ventral (V), lateral (L), and medial (M) directions. **B**, UMAP visualizations of the samples from two subjects (P and F). **C**, UMAP visualizations of striatal nuclei colored by the three regions. The color scheme for these regions is the same as in **A**. **D**, Feature plots of canonical neuronal and astrocyte marker gene expression in striatal nuclei. **E**, UMAP visualization colored according to eight major classes in the NHP. **F**, Heatmap of differentially expressed genes. Color bar at the top corresponds to the major classes identified in **E**. **G**, Violin plots of distributions of marker gene expression across nine clusters, with MSNs divided into D1- and D2-MSNs.

on their gene expression counts (Methods). Each major cluster had nuclei derived from both subjects (Figure 17B) and all regions (Figure 17C), indicating broad similarities between the subjects and the regions. Feature plots showed the expression of well-known marker genes and thus indicated the correspondence between nuclei clusters and broad cell classes including D1-MSNs, D2-MSN, striatal interneurons, and

non-neuronal cell types (Figures 17D, S1C and S2). Each major cell class was signified by groups of differentially expressed genes that were specifically enriched in that cluster (Figures 17E and 17F; Data S1). The intersection of differentially expressed genes from each cluster with orthologous marker gene lists from several databases[286,290,303] confirmed the identity of the clusters ($p < 0.05$, Benjamini-Hochberg corrected, Hypergeometric tests). Violin plots showing the expression levels of marker genes in each major class confirmed the basic validity of our experimental and analytic methods (Figure 17G). Together, these results provide a broad transcriptional catalogue of cell classes in the NHP striatum and a rich dataset for future exploration.

4.2.2 **Transcriptional Diversity of Medium Spiny Neurons**

The expression of well-known MSN marker genes, including *PP1R1B*, *BCL11B*, and *PDE1B,*[316,317] indicated which clusters contained MSNs (Figures S1E-S1G). Differential gene analysis of those MSN clusters vs. other clusters revealed several other MSN markers, including *KIAA1211L*, *PDE2A, SLIT3*, and *NGEF* (Figures S1H-S1K). To identify MSN subtypes, we recalculated the PCAs and performed dimensionality reduction on the isolated MSN nuclei, clustered them at a resolution that distinguished physically separated UMAP clusters, and annotated them using FISH probes against a mixture of previously described[289,318,319] and novel marker genes (Data S1; Figures 18A-C and 20-23). The clusters putatively corresponded to D1- and D2-MSNs in the matrix (D1M and D2M), D1- and D2-MSNs in striosome (D1S and D2S), D1- and D2-MSNs in the NAc shell and olfactory tubercle (OT) (D1Sh and D2Sh), and MSN-like neurons located in the interface islands (Figure 18A). One cluster was a D1/D2-hybrid (D1/2) and shared many characteristics with a novel MSN type described in rodents (D1H or eccentric-

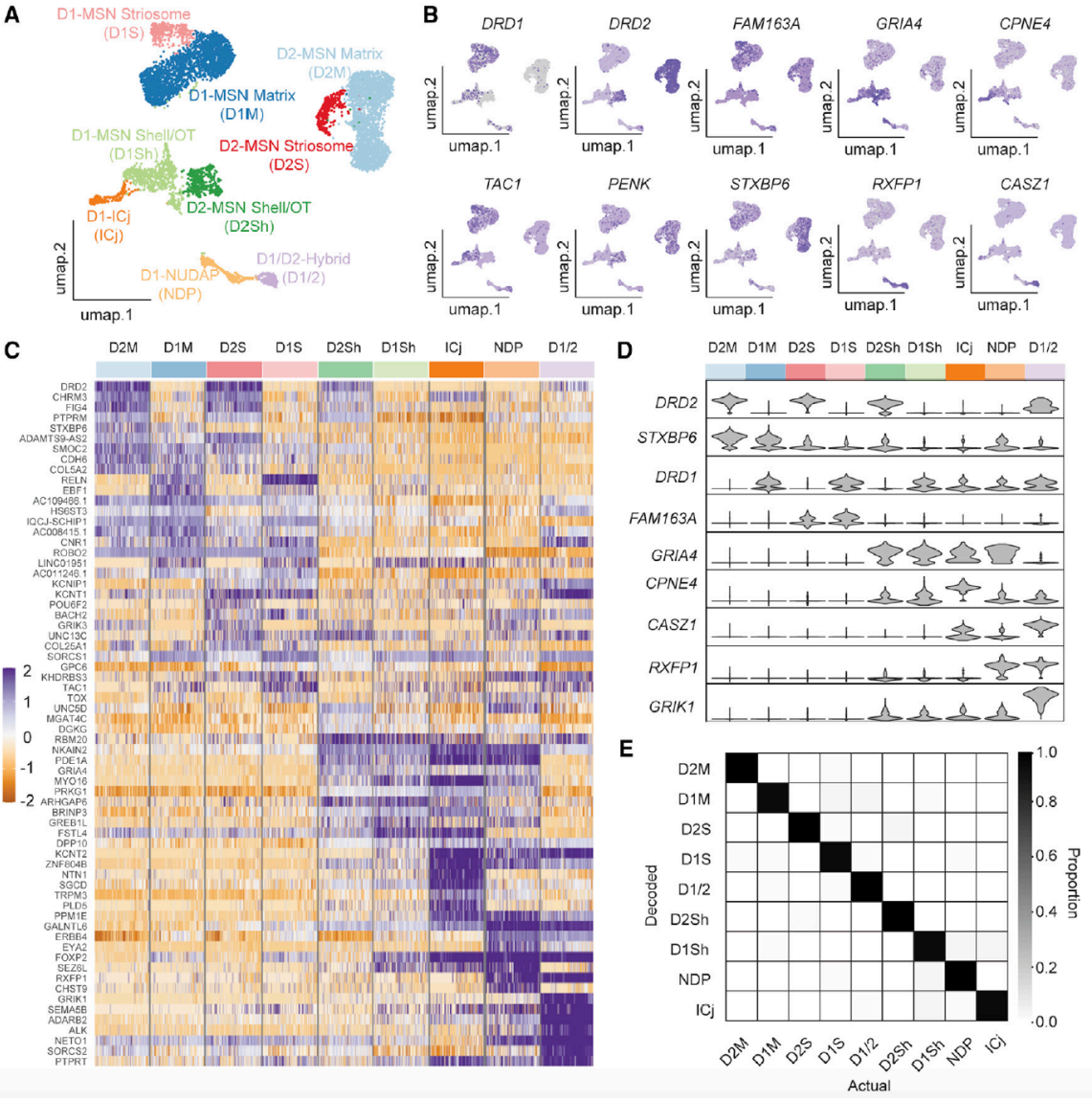


Figure 18: Medium spiny neuron (MSN) subtypes

A, UMAP projection of MSN nuclei. Each dot represents a nucleus, and the colors represent the different MSN types. **B**, Feature plots for the expression of *DRD1*, *TAC1*, *DRD2*, and *PENK* showing the separation of D1- and D2-MSNs and the expression of marker genes enriched in each cluster. **C**, Heatmap showing the top ten most enriched genes in each MSN type. Colored bar at the top corresponds to the colors in **A**. **D**, Top: MSN-type identifications colored according to **A**. Bottom: violin plots showing cell-type- and compartment-specific marker gene expression. **E**, The accuracy rate between SCCAF-decoded cell type and actual cell type using the data combined from both subjects.

SPN)[288,290]. Each MSN cluster was signified by groups of differentially expressed genes that were specifically enriched in that cluster (Figures 18C and 18D; Data S1; Table S1) (Methods). We used the Single Cell Clustering Assessment Framework (SCCAF) to quantify the robustness the MSN subtype clusters[312]. For all nine clusters, the model predictions were robust: the areas under

the receiver operating characteristic (AUROC) was more than 0.92 and 0.94, for monkeys P and F, respectively (Figure 18E and S3D)[312]. Together, these results indicate that the primate striatum contains at least nine transcriptionally distinct neuron subtypes that all feature characteristics of MSNs.

We used archetypal analysis to determine the similarity between the monkeys P and F, and the relationships between MSN subtypes[320-322]. Archetypal analysis decomposes the expression matrix into gene loading vectors, or 'archetypes,' that correspond to cell states[320-322]. We found the archetypes that defined the transition from one MSN subtype to another using the data from Monkey P. When projected onto monkey F, these archetypes maintained the same transitions between subtype clusters (Figure S3G). This result illustrates the high level of consistency between the two experimental subjects and validates the archetypes. We subsequently found all genes that were significantly correlated with the archetypes that defined the transitions between subtypes ($p < 0.05$, Benjamini-Hochberg corrected, Pearson's correlation). Each subtype pair contained genes where transitions included a significant discontinuity (Figure 19A, $p < 0.05$, Regression discontinuity test, Methods) and other genes without significant discontinuity (Figure 19B, $p > 0.05$, Regression discontinuity test). The *p*-value distributions demonstrated the degree of discontinuity between subtype pairs (Figure 19C). For every pair, except for D1S and D2S – the subtypes for which we had the lowest sample sizes – the *p*-value distributions indicated more statistically significant discontinuities than predicted by chance (Figure 19C, $p < 0.00001$, Benjamini-Hochberg corrected, Binomial test, Methods). These results demonstrate that each identified MSN subtype is characterized by discrete gene expression patterns.

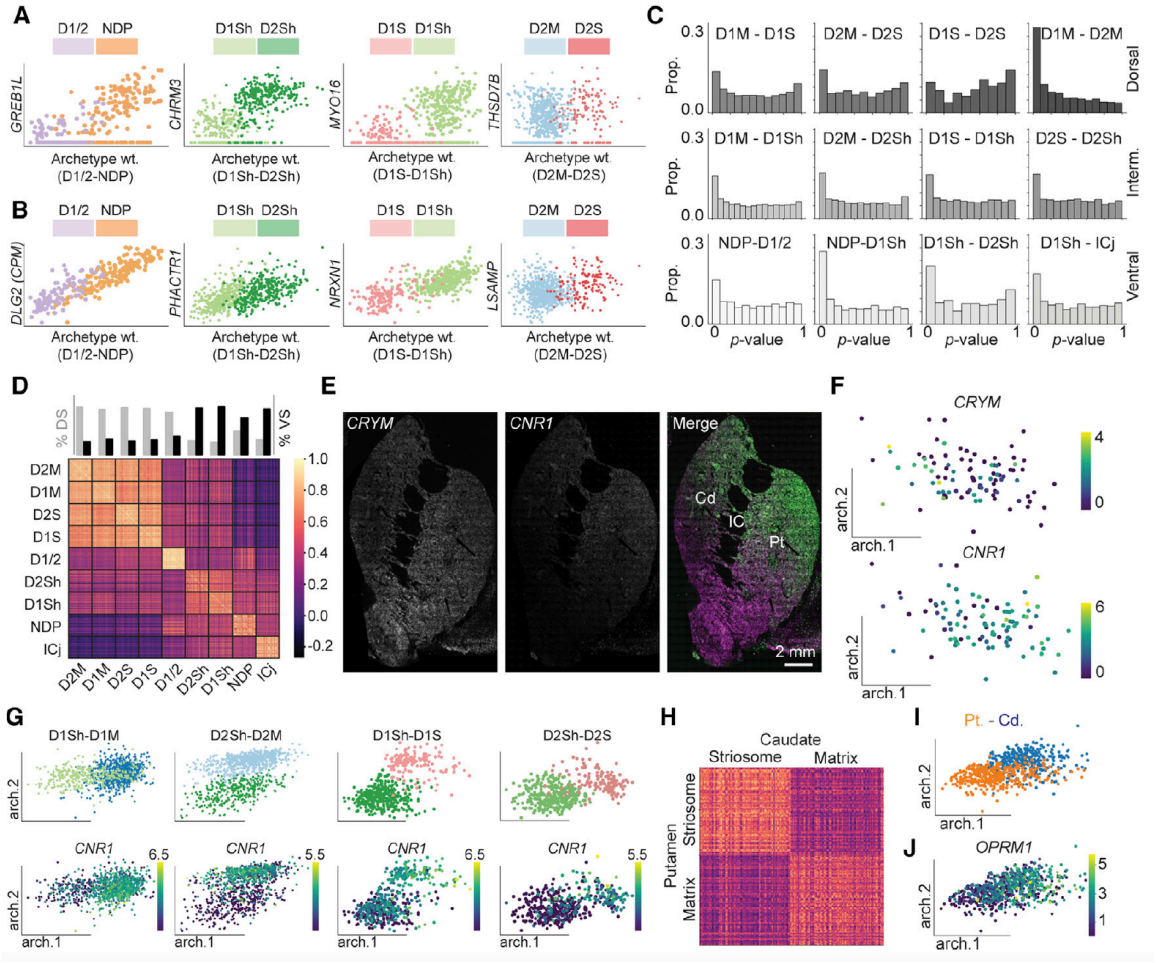


Figure 19: Archetypal analysis of MSN subtypes
A, Representative genes showing significant discontinuity between a subtype pair. **B**, Representative genes showing non-significant discontinuity between a subtype pair. **C**, The *p* value distributions between each subtype pair. **D**, Heatmap showing the cosine similarity between cells within and between the nine types of MSNs. **E**, FISH labeling of *CRYM* (magenta) and *CNR1* (green) reveals a continuous gradient on the dorsal-ventral axis. **F**, *CRYM* (top) and *CNR1* (bottom) expressions along the D1/D2-hybrid archetype axes. **G**, The archetype distribution of subtype pairs (top) that were divided between the DS and VS and *CNR1* (bottom) distributions in these archetype axes. **H**, Heatmap showing the cosine similarity between cells within and between striosome and matrix in Cd and Pt. The color scale is the same as in **D**. **I**, Distribution of Cd and Pt in archetype axes. **J**, *OPRM1* expression in the Cd and Pt cells along the archetype axes.

Despite our emphasis on discrete borders between subtypes, we also detected continuous gradients of gene expression. Cosine similarity between nuclei indicated that DS-derived clusters were more like other DS clusters compared to VS-derived clusters, and vice versa (Figure 19D) ($p < 0.0001$, Permutation tests, STAR Methods). Consistent with this and with previous studies in mouse,[288,289] the crystallin mu and the cannabinoid receptor genes, *CRYM* and *CNR1*, respectively,

reflect continuous gradients on the dorsal-ventral axis of the mouse striatum (Figure 19E). Archetypal analysis indicated that this gradient is reflected in the D1/D2-hybrid population (Figure 19F; Table S2), and across subtype pairs that were divided between the DS and VS (Figure 19G). This result highlights that gene expression gradients that define position on the dorsal and ventral axis are conserved between species. The primate DS is divided into the Cd and Pt by the internal capsule. Compared to the differences between the DS and VS, the Cd and Pt appeared more similar. Indeed, cosine similarities between the striosome and matrix showed that striosome nuclei in the Cd were far more similar to striosome nuclei in the Pt, as opposed to matrix nuclei in the Cd (Figure 17H, $p < 0.0001$, Permutation test, STAR Methods). Nevertheless, archetypal analysis identified a gradient that showed an enhanced μ-opioid receptor (*OPRM1*) expression in the Cd (Figures 19I and 19J; Data S2). This archetype might highlight striosome-enriched nature of the Cd.

Archetype analysis also highlighted continuous sources of variation within subtypes. Some nuclei derived from the VS clustered together with matrix nuclei (D1M and D2M). We detected archetypes that highlighted the VS nuclei fraction (Figure S3H), and the genes that were correlated with the VS weighted vector are preferentially expressed in NAc core, including *ZDBF2* ($r = 0.18$ and 0.2, $p < 10^{-6}$ and 10^{-7}, with D1- and D2-MSNs, respectively, Pearson's correlation) and *HPCAL4* ($r = 0.12, p < 10^{-3}$ with D2-MSNs, Pearson's correlation) (Data S3)[323]. Another archetype of particular interest was found in D1Sh and defined by upregulation of the gene for the Neurokinin-B receptor, *TAC3*. Differential gene analysis of the D1Sh *TAC3*-archetype revealed selectively genes enriched in this archetype, including *MPPED1*, *HPCAL1*, and *MEIS3*. This result demonstrates within-subtype variations with potentially functional roles.

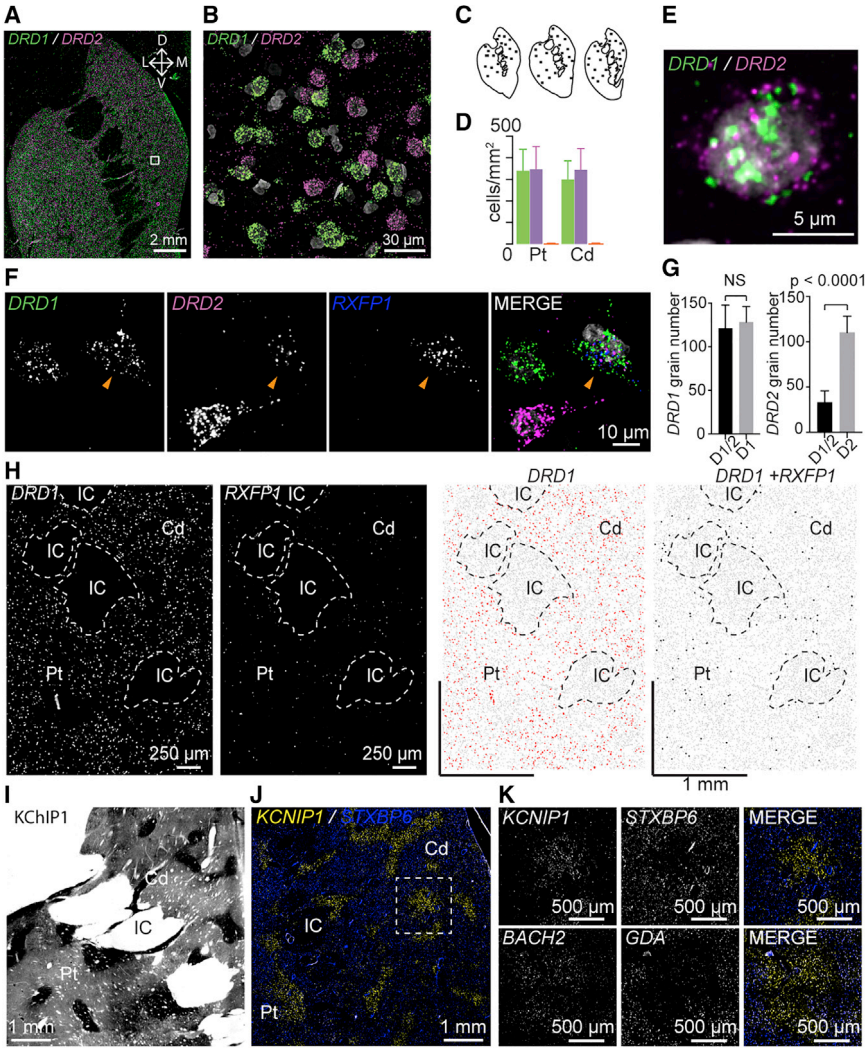


Figure 20: MSN subtypes in the dorsal striatum

A, FISH labeling of *DRD1* (green) and *DRD2* (magenta). White box indicates the region shown in **B**. The top right axis shows D, V, L, and M directions. **B**, High-resolution image of region highlighted in **A**. **C**, Schematic diagrams of the three sections used for *DRD1* and *DRD2* quantification. The square boxes indicate the quantified regions of interest (ROIs). **D**, Quantification of cell density of neurons expressing *DRD1* (green), *DRD2* (magenta), or both (orange) in the Cd and Pt. Error bars are SD across ROIs. **E**, One example MSN expressing both *DRD1* and *DRD2*. **F**, *RXFP1* labels D1/D2-hybrid MSNs in the dorsal striatum. Arrowhead points to an example D1/D2-hybrid cell. **G**, Quantification of *DRD1* and *DRD2* grain number in D1/D2-hybrid cells and normal D1- or D2-MSNs. Unpaired t-test was used for statistical analysis, and *p* values are indicated on the plots. Error bars represent standard deviation (SD) across 6 cells per type. NS, non-specific. **H**, FISH labeling of *DRD1* and *RXFP1*. Left two pictures are original FISH images showing the distribution of *DRD1* and *RXFP1*. The right two pictures are CellProfiler-processed images showing *DRD1* expressing (red dots) or *DRD1* and *RXFP1* (black dots, enlarged for display purposes) co-expressing cells. Gray dots are nuclei. **I**, Immunohistochemistry of KChIP1 showing robust striosome pattern. **J**, FISH labeling of *KCNIP1* (yellow) and *STXBP6* (blue) distinguishes striosome and matrix, respectively. White square indicated the region shown in **K**. K, Top: detail of the white square in (J). Bottom: as above, for characteristic striosome and matrix markers, *BACH2* and *GDA*.
Cd, caudate; Pt, putamen; IC, internal capsule.

4.2.3 **Medium Spiny Neuron Subtype and Archetype Distributions in the Dorsal Striatum**

We used FISH to explore the anatomical distribution of MSN subtypes and archetypes. As expected, the majority of neurons expressed either *DRD1* (325 ± 99 nuclei/mm^2) or *DRD2* (320 ± 99 nuclei/mm^2; Figures 20A-D). To investigate the relatively rare hybrid cell type (D1/2) that expresses both *DRD1* and *DRD2*, we used probes against *RXFP1*, a highly specific

marker gene for D1/2 cells ($p = 2.47 \times 10^{-135}$, Wilcoxon). High resolution imaging confirmed that *RXFP1* positive cells in the dorsal striatum co-labeled with *DRD1* and *DRD2* (Figure 20F). D1/2 cells were distributed uniformly throughout the DS (Figure 20G). Intestingly, D1/2 cells had the same amount of *DRD1* expression as nearby D1-MSNs, but there was far less *DRD2* expression when compared to nearby D2-MSNs (Figure 20H, $p = 0.55$ and $p < 0.0001$ respectively, unpaired *t*-tests). Co-clustering of the NHP data with striatum data from mouse indicates that the D1/2 shares important transcriptional characteristics with a subset of the recently described "D1H" (Figures S3A-S3C)[288,290]. These results confirm that a hybrid D1/2 cell type is also present in the NHP striatum.

Both D1- and D2-MSNs derived from the DS split into two clusters; one larger cluster and one smaller cluster. We reasoned that the larger D1- and D2-MSN clusters likely corresponded to the matrix (Figure 18A, dark and light blue clusters, respectively), whereas the smaller clusters likely corresponded to striosomes (Figure 18A, light and dark red clusters, respectively). The genes most enriched in the striosome MSNs (D1S and D2S) were *KCNT1*, *KHDRBS3*, *FAM163A*, *BACH2*, and *KCNIP1*, whereas the genes most enriched in the matrix MSNs (D1M and D2M) were *EPHA4, GDA*, *STXBP6*, and *SEMA3E* (Figures 2C and S4A) ($p < 1.77 \times 10^{-52}$, Wilcoxon). *KCNIP1* is the gene for the the potassium channel-related KChIP1, and staining for KChIP1 labels the boarders of the striosome (Figure 20I)[324]. We performed FISH labeling with *KCNIP1* and *STXBP6* – a highly specific matrix marker – and the results show a similar pattern as the KChIP1 antibody labelling (Figures 20J and 20K). Similar patterns were observed using FISH probes that labeled other identified striosome and matrix markers (Figures 20K and S4B-S4D). Thus, these results confirm that four major snRNA-Seq clusters derived mostly from the DS correspond to D1- and D2- MSNs in the striosome and matrix.

4.2.4 **Medium Spiny Neuron Subtype and Archetype Distributions in the Ventral Striatum**

Four MSN clusters were highly enriched in the nuclei from the VS (Figure 19D). Three of the clusters were *DRD1* positive, whereas the fourth cluster was *DRD2* positive. Differential gene analysis of the two largest clusters (Figure 18A, light and dark green) revealed selectively enriched marker genes, including *GREB1L*, *ARHGAP6*, and *GRIA4* (Figure 18C). FISH labeling of the striatum with *DRD1* and *ARHGAP6* revealed that *ARHGAP6* was enriched in a restricted portion of the VS that likely corresponds to the NAc shell and OT (Figures 21A, 21B and S5A), but was not prevalent in the core (Figure S5B). Quantification of grain number of *ARHGAP6* in putative shell/OT and core regions confirmed that *ARHGAP6* was more enriched in shell/OT (Figure 21B, left), and very similar pattern was observed with probes against another enriched marker gene, *GREB1L* (Figure 21B, right). We labeled sections with *DRD1* and *GREB1L* probes and an adjacent section with an immunofluorescent stain for calbindin, which roughly markes the border between NAc core and shell[325]. The *GREB1L* intensity traced the putative transition from the calbindin-poor shell to the calbinin-rich core (Figure S5C). Thus, these results indicate that NAc Shell and OT are comprised of region-specific D1- and D2-MSNs.

Archetypal analysis revealed a *TAC3*-positive archetype within the D1-Shell/OT subtype (Figures 21C and S5D-S5F; Table S3). In order to locate this MSN archetype and to distinguish it from *TAC3*-positive interneuons (Figure S2)[291], we triple labelled coronal sections with probes against *DRD1*, *DRD2*, and *TAC3* (Figures 21D and 21E). This labeling revealed the previously described *TAC3* interneuons unifomly distributed throughout the striatum (Figure 21F); these cells did not show co-localization with *DRD1* or *DRD2*. However, there was a cluster of *TAC3*-positive neurons located in the medial shell region of the NAc that co-localized with *DRD1*, and far fewer

than that co-localized with *DRD2* (Figures 21G and 21H). We observed the same *DRD1*- and *TAC3*-positive cluster in the medial shell region of a second monkey (Figure 21I). These results reveal a novel *TAC3*-positive MSN archetype in the primate NAc.

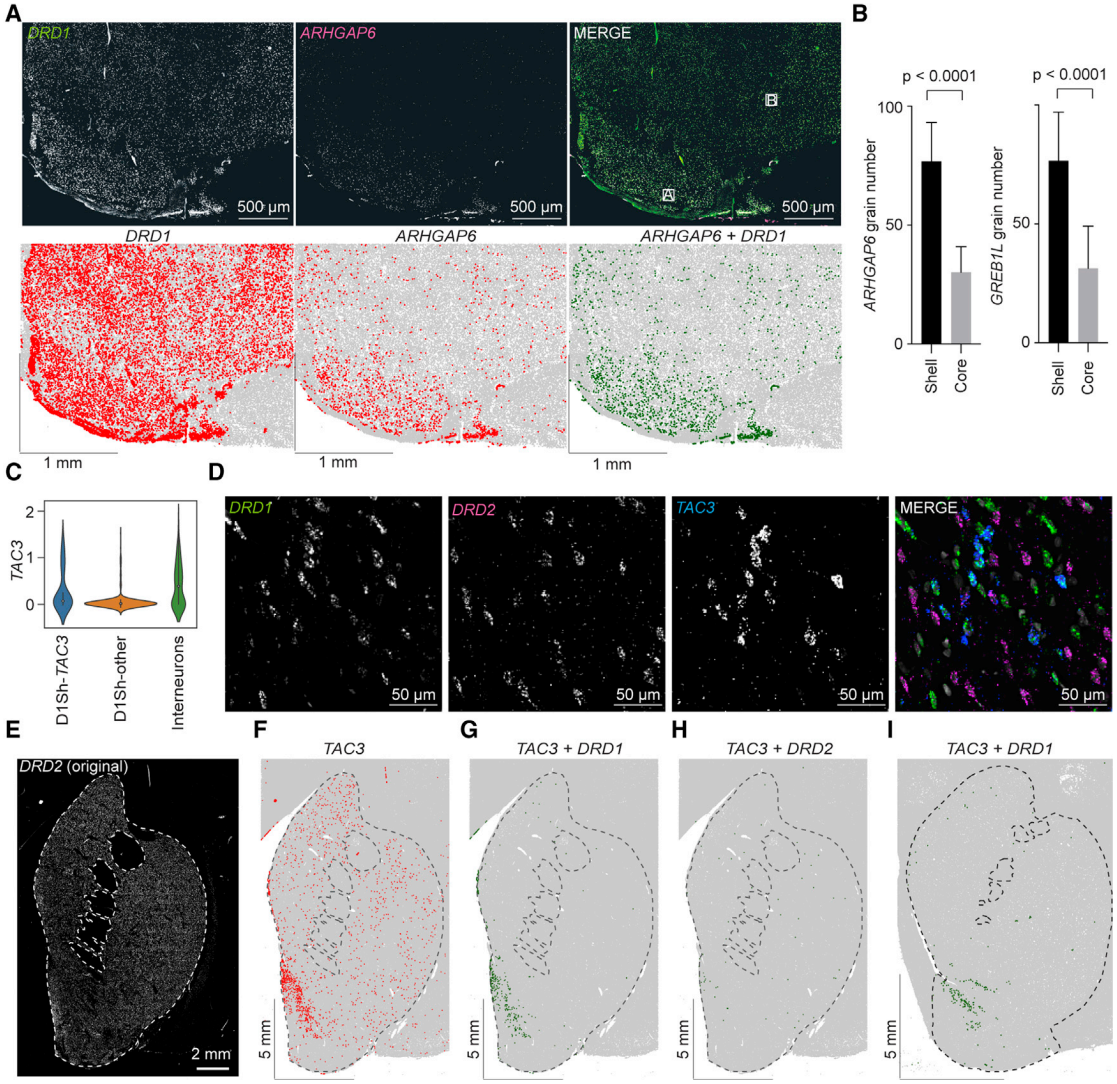


Figure 21: MSN subtypes in the VS

A, Top: double labeling with *DRD1* (left) and *ARHGAP6* (middle) of the NAs shell/OT shows that *ARHGAP6* is selectively enriched in the shell/OT. Bottom: CellProfiler-processed images for the above images. Lettered boxes indicate regions shown in Figure S5A and S5B. **B**, Left: quantification of grain number of *ARHGAP6* in shell/OT and core. Unpaired t test was used for statistical analysis. Error bars represent SD across thirty-two cells per each region. Right: quantification of grain number of *GREB1L* in shell/OT and core. Unpaired t test was used for statistical analysis. Error bars represent SD across twenty-nine cells per each region. **C**, Violin plot showing *TAC3* levels in D1Sh-*TAC3* archetype, other D1Sh cells, and *TAC3* interneurons. **D**, TAC3 co-localizes with *DRD1* in medial shell MSNs. **E**, *DRD2* FISH image showing the outline of the striatum. Dashed white line delineate the borders of striatum. **F**, CellProfiler results showing the *TAC3* distribution (red dots) in the section in **E**. Gray dots are nuclei. **G**, CellProfiler results showing the distribution of *TAC3* and *DRD1* co-expressing cells (dark green dots) in the section in **E**.
(H) CellProfiler results showing the distribution of TAC3 and DRD2 co-expressing cells (dark green

One of the most remarkable features of the *DRD1* and *DRD2* labelling in the VS was the presence of D1-exclusive islands that likely corresponded to "interface islands" (Figure 22A)[326,327]. We investigated whether the cell types within these D1-exclusive islands corresponded to the smaller *DRD1* enriched VS clusters (Figure 18A, light and dark orange). We selected two

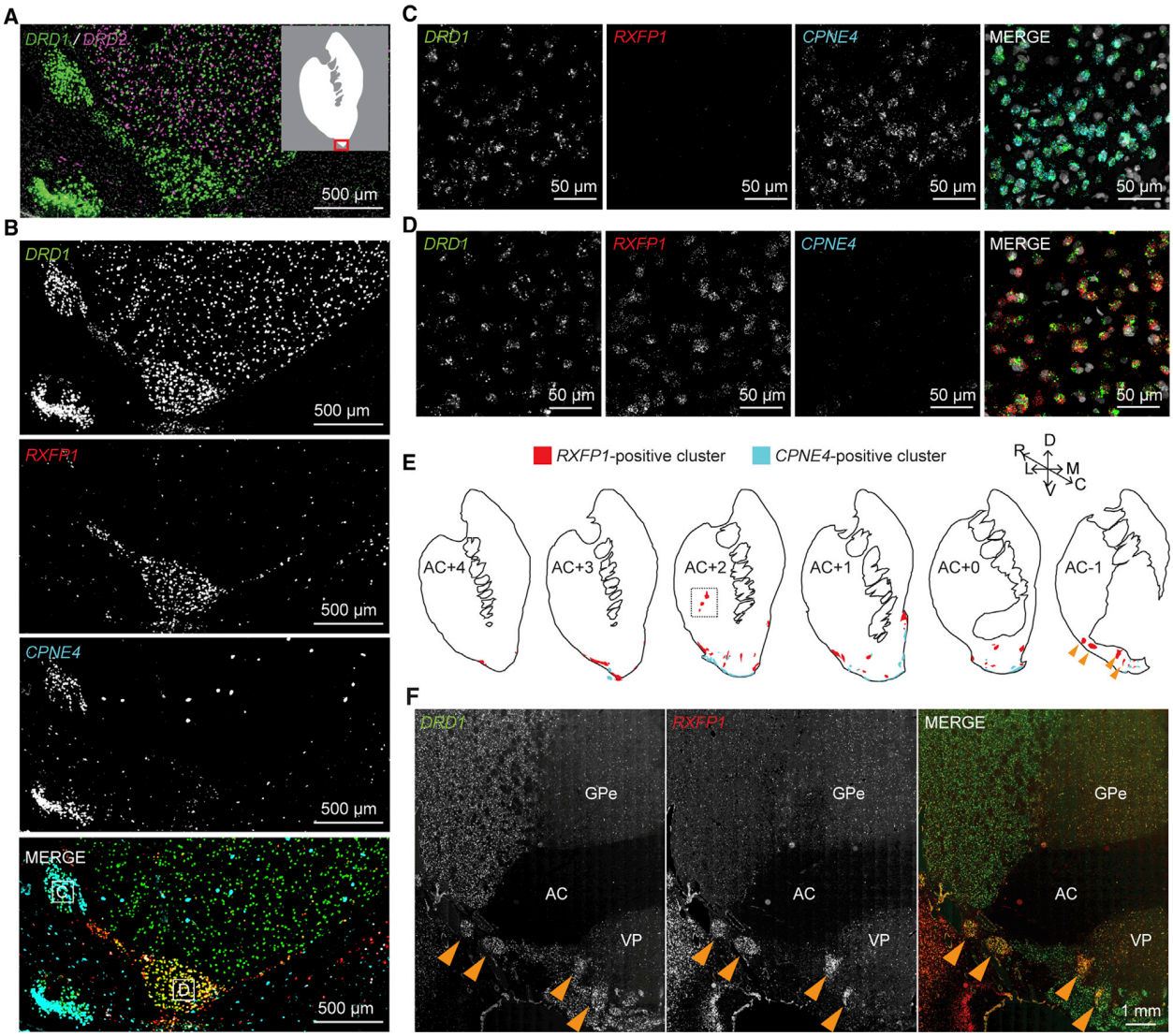


Figure 22: Cell types in the interface islands
A, FISH stain of *DRD1* (green) and *DRD2* (magenta). *Inset*: White area indicates striatum and the red box highlights the area shown in **A** and **B**. B, FISH stain of *DRD1*, *RXFP1* and *CPNE4* in immediately adjacent section from **A**. **C**, High-resolution image of the regions indicated with the letter "C" in **B**. D, High-resolution image of the region indicated with the letter "D" in **B**. **E**, Distribution of *RXFP1* and *CPNE4* clusters across six rostral-caudal regions identified by multichannel FISH. The upper right axis shows dorsal (D), ventral (V), lateral (L), medial (M), rostral (R), and caudal (C) directions. Dashed white box denotes *RXFP1* clusters in putamen, the FISH image of which is shown in Figure S5H. Yellow arrowheads point to *RXFP1* clusters in caudal extent of NAc in both illustration and images shown in **F**. **F**, Example *RXFP1* clusters in caudal extent of NAc. AC = anterior commissure, GPe = external globus pallidus, VP = ventral pallidum.

respective marker genes, *RXFP1* and *CPNE4*, for the two *DRD1* enriched VS clusters (Figure S6C). We labeled one section using *DRD1*, *RXFP1,* and *CPNE4* probes and revealed that *RXFP1* and *CPNE4* labeled distinct D1-exclusive cell islands (Figure 22B). High resolution confocal microscopy of these islands verified that *CPNE4* and *DRD1* co-localized in the same cells in one island (Figure 22C), whereas *RXFP1* and *DRD1* co-localized in the same cells of another island (Figure 22D). These results suggest that different interface islands contain different *DRD1*-positive cell types.

We repeated the *DRD1*, *RXFP1*, and *CPNE4* FISH on regularly spaced pre- and peri-commissural coronal sections. We defined the regions of the VS by comparison of Nissl-stained sections with a high-resolution MRI and DTI Rhesus macaque brain atlas (Figures S6A and S6B)[328,329] and we mapped all the nearby islands in two monkeys (Figures 22E and S5G-S5I). The *CPNE4*-positive islands appeared to correspond to the Islands of Calleja (ICj)[330]. This correspondence was verified by intense co-localization of *CPNE4* and *DRD1* in the cells of the major ICj, an easily identifiable landmark at the border between the NAc and the septal nuclei (Figure S6D). Previous studies have shown that cells in ICjs are granule cells[330]. Likewise, the *CPNE4*-positive cells we examined were small and exhibited high packing density (Figures S6E and S6F). Thus, the cluster enriched with *DRD1* and *CPNE4* corresponded with granule cells in the ICjs. Outside of the ICjs, co-localization between *DRD1* and *CPNE4* was restricted to a dense cell layer at the ventral extreme of the OT, possibly corresponding to a portion of the anterior olfactory nucleus (AON) (Figures 22B and S6G)**.** Gene enrichment analysis revealed that differentially expressed genes in these cells are implicated in neurogenesis, neurosecretion, and many other functions (Data S4).

In contrast to the *CPNE4*-positive ICjs, the *RXFP1*-positive islands had larger nuclei that were less densely packed together and did not appear different from nearby D1-MSNs (Figures S6E and S6F). Likewise, *RXFP1*-positive islands were not restricted to the border regions of the VS, rather, they were found throughout the NAc, putamen, and near the adjacent septal nuclei (Figure 22E, orange arrows and dashed black box, Figures S5G-S5I, S6H, S6I, and S7A). *RXFP1*-positive cells located in these VS islands exhibited high levels of *DRD1* expression, but no detectable *DRD2* expression (Figures S7A and S7B). Moreover, cells in the *RXFP1*-positive islands expressed high levels of the gene for the μ-opioid receptor (*OPRM1*) compared to regular D1- MSNs located outside of the D1-exclusive island (Figures 23A and 23B). Interestingly, expression of the κ-opioid receptor gene (*OPRK1*) was almost completely absent from these islands, compared to surrounding tissues (Figure S7D). Using immunohistochemistry, we confirmed that MOR was expressed in the *RXFP1*-positive islands (Figure 23C). Based on their distribution and the upregulation of μ-opioid receptor – upregulation that was not present in the ICjs (Figures 23D-F) – we concluded that these *RXFP1*-positive interface islands corresponded with Neurochemically Unique Domains in the Accumbens and Putamen (NUDAPs)[283,284]. Therefore, we denoted the *DRD1*- and *RXFP1*-positive cells as D1-NUDAP neurons. Gene enrichment analysis revealed that D1-NUDAP neurons express genes that have been implicated in drug addiction and many other functions (Data S4). These data show a novel cell type that is associated with interface islands and could be critical for the hedonic aspects of reward. Altogether, these results demonstrate that the ventral striatum is characterized by the presence of discrete MSN subtypes that correspond to functionally relevant subdivisions including NAc shell, OT, and distinct types of interface islands.

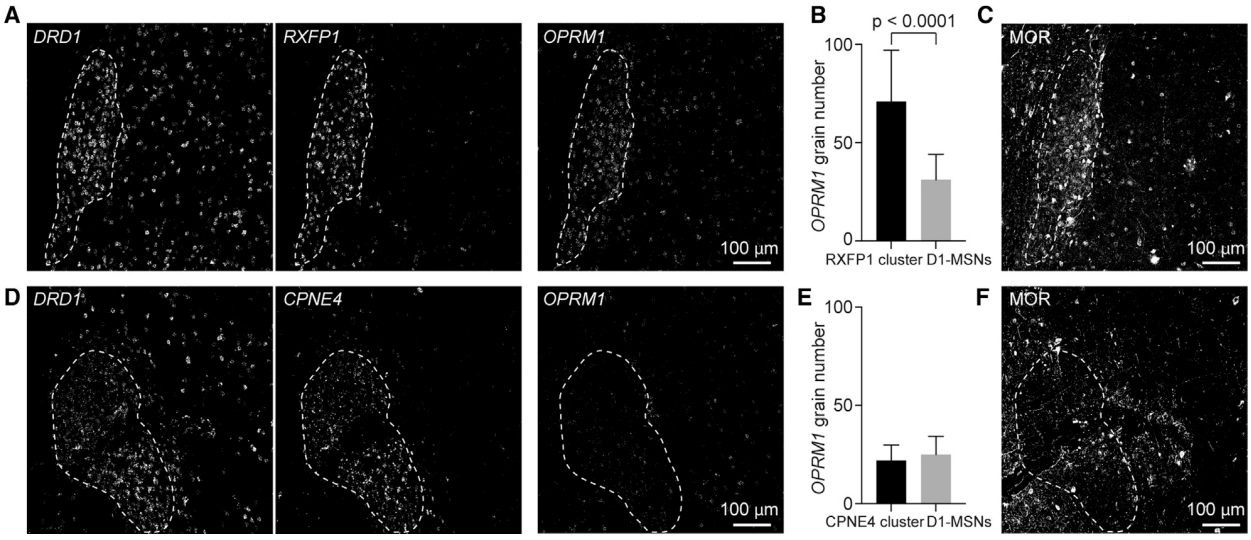


Figure 23: μ-opioid receptor expression is specifically enriched in D1-NUDAP cells

A, FISH stain of *DRD1* and *RXFP1* as well as *OPRM1* in two close sections. White dashed line delineates the boundaries of a *RXFP1* cluster. **B**, Quantification of grain number of *OPRM1* in *RXFP1* clusters and close D1-MSNs. Unpaired t-test was used for statistical analysis. Error bars represent standard deviation (SD) across 41 cells in each group. **C**, MOR expression in an adjacent section of **A**. **D**, FISH stain of *DRD1* and *CPNE4* as well as *OPRM1* in two close sections. White dashed line delineates the boundaries of a *CPNE4* cluster. **E**, Quantification of grain number of *OPRM1* in *CPNE4* clusters and close D1-MSNs. Unpaired t-test was used for statistical analysis. Error bars represent standard deviation (SD) across 36 cells in each group. **F**, MOR expression in an adjacent section of **D**.

4.3 **Conclusion**

The phylogenetic relationship between humans and nonhuman primates (NHPs) makes NHPs a crucial neuroscientific model. Here, single nucleus RNA-sequencing (snRNA-Seq) revealed at least nine distinct Medium Spiny Neuron (MSN) and MSN-like subtypes in the NHP striatum (Figure 17). The borders between subtype pairs were characterized by discontinuities in gene expression, though we also found continuous axes of variation (Figure 18). We identified five distinct MSN subtypes in the dorsal striatum (DS), including D1- and D2-MSN subtypes specific to the striosome and matrix compartments, as well as a hybrid cell type that contained mRNA for both D1 and D2 receptors (Figure 20). The ventral striatum (VS) contained at least four

distinct subtypes, including D1- and D2-MSN subtypes located specifically in the nucleus accumbens (NAc) shell and olfactory tubercule (OT) regions (Figure 21), and two subtypes associated with the "interface islands" – dense cellular islands located within and near the ventral border of the striatum. Marker genes for one of these VS cluster subtype were highly enriched in the Islands of Calleja (ICjs) (Figure 22). Cells from the other VS cluster subtype were restricted to Neurochemically Unique Domains in the Accumbens and Putamen (NUDAPs) (Figures 22 and 23)[283,284]. Within these subtypes, archetypal analysis revealed finer distinctions, including a *TAC3*-positive D1Sh-MSN that could represent the origin of a third pathway through cortico-basal ganglia loops (Figure 21)[331]. Together, these MSN subtypes and archetypes provide a blueprint for studying cell-type specific functions during sophisticated primate behaviors, and the cell-type-specific marker genes define potential molecular access points to enable the application of genetically coded tools in scientific or translational contexts.

The DS is divided into the caudate nucleus (Cd) and putamen (Pt) by the internal capsule. Spanning these structures are two neurochemical compartments, matrix and striosome, that form the 'neostriatal mosaic'[278,280]. Broadly understood, matrix MSNs receive neocortical inputs from associative and sensorimotor cortex and give rise to the direct and indirect pathways[332]. In contrast, striosomal MSNs, and the recently discovered, extrastriosomal 'exo-patch' MSNs[289,319], receive input from limbic territories, including the anterior cingulate cortex, orbitofrontal cortex, and anterior insular cortex[333], and project directly to midbrain dopamine neurons[334,335]. Despite our advanced understanding of these circuit-based structures, we are only beginning to gain insights into the associated circuit-based functions. For example, striosomal MSN activations influence cognitive and emotional decision making[336] and value based learning[337]. A critical milestone that will enable us to accelerate functional discovery will be the development of cell-type- and

compartment-specific viral vectors to enable circuit interrogation in NHPs. Here, we identified compartment-specific gene expression patterns for NHP matrix – *STXBP6*, *GDA*, and *SEMA3E* – and striosome – *BACH2*, *KCNT1*, *KCNIP1*, and *KHDRBS3*. Moreover, we found an extrastriosomal cell type, the D1/D2 hybrid, that expressed many of the genes associated with striosome, thus suggesting a possible homology to 'exo-patch' cells[319]. We expect that understanding the regulatory vocabulary governing these gene expression patterns will reveal cell-type-specific enhancers that can be packaged into AAVs that grant molecular access to striosome and matrix MSNs in the NHP brain.

The VS is strongly implicated in reward processing[282]. The NAc and OT complex comprises a major portion of the rostral VS, and this complex is traditionally recognized as a limbic-motor interface[338]. The NAc is further divided into core and shell territories with distinct behavioral functions[339]. We found several gene markers, including *ARHGAP6*, *GREB1L*, and *GRIA4,* that were upregulated in the VS samples (Figure 18). FISH labeling with these probes revealed that their upregulation traced the transition of the ventrally positioned, Calbindin-poor shell to the dorsally positioned, Calbindin-rich core (Figure S5C). Thus, these marker genes label D1- and D2-MSN subtypes that were specific to the NAc shell and OT. Within the D1Sh subtype, we detected an archetype that expresses the gene for Neurokinin-B (*TAC3*) (Figures 21C-21I). Previous tracing studies in rodents have shown that a small population Neurokinin B-positive D1-MSNs have direct projections to other basal forebrain structures, including notably the cholinergic *substantia inomiata*[331,340,341]. Thus, this *TAC3* MSN archetype could represent the genesis of an additional pathway, along with the direct and indirect pathways, for cortico-striatal signals to reach the cortex. As with the DS cell types, the regulatory code controlling these cell-type-specific VS gene expression patterns will likely hold the keys to molecular access points.

The DS striosome and the VS are both implicated in limbic functions and reward processing, and thus it is interesting to compare these subpopulations. Our data indicate that there are many transcriptional similarities between the striosome and VS cell types, but also some key differences. For example, many striosome specific markers are upregulated in D1-NUDAP cells, including *KCNIP1*, *KCNT1*, *KHDRBS3*, and *BACH2* (Figure S7C). Even *PDYN*, which is a widely acknowledged D1-striosome marker gene[342], is also expressed in D1-NUDAPs. On the other hand, D1-NUDAPs also express some genes which we found to be selectively expressed in the matrix, including *STXBP6*, *GDA*, and *SEMA3E*. *OPRM1* was upregulated in the striosome, as predicted, but the upregulation was not as dramatic as we expected. In contrast, we observed robust *OPRM1* signals in the NUDAPs (Figures 23A-23C). Indeed, this selective enrichment of *OPRM1* in the *RXFP1*-positive interface islands is a key piece of evidence in favor of the NUDAP hypothesis. This selective *OPRM1* enrichment suggests the intriguing possibility that NUDAPs are part of the network of "hedonic hotspots?" Hedonic hotspots are regions in the NAc and ventral pallidum that, when opioids are directly applied, produce behavioral reactions that indicate pleasure[343]. As might be expected for hedonic hotspots, the κ-opioid receptor gene (*OPRK1*) was absent from D1-NUDAP cells (Figure S7D). These differentially expressed genes and others provide a blueprint for understanding cell-type-specific contributions of this novel cell type to reward processing and pleasure.

The basal ganglia are highly conserved throughout vertebrate evolution[344], and a rough comparison of our results with single cell studies of the mouse striatum[286-290] confirmed this pattern at the level of cell types. As with prior studies, we find a relatively even distribution of D1- and D2-MSN in the striatum. Approximately 10% of sampled MSNs were identified as striosome MSNs[290]. We found that our neuronal nuclei samples contained approximately 85% MSNs and

15% interneurons; this is a higher proportion of interneurons than was recovered from single cell analysis of rodent striatum[290], but consistent with counts made in humans[345]. A novel and relatively rare MSN type has been recently documented and variously described as eccentric SPNs[289,290], *Pcdh8*-MSNs[286], and D1H[288]. Both D1/D2 hybrid and D1-NUDAP neurons shared some characteristics with this novel cell type. We formally compared our results to D1H because the sequencing depth was most similar between the studies. Co-clustering our data with the mouse data revealed that approximately half of the D1H population co-clustered with D1/D2-hybrids, and the other half co-clustered with D1-NUDAPs (Figures S3A-S3C). However, despite their similarities and their co-clustering with D1H, our data indicate that D1-NUDAP neurons and D1/D2-hybrid neurons represent distinct MSN subtypes. First, although *DRD2* expression was common in D1/D2-hybrids (Figures 18D and 20F), we found no evidence of *DRD2* expression in D1-NUDAPs (Figures S7A and S7B). Second, the D1-NUDAP neurons did not express other marker genes, including *CASZ1* and *GRIK1,* that were reported in D1/D2-hybrids, D1H, and eSPNs[288-290]. Third, a machine learning classifier easily distinguished between D1-NUDAP and D1/D2-hybrid neuron subtypes (Figure 18E). Finally, we only found D1/D2-hybrid MSNs in the DS, whereas D1-NUDAP cells were restricted to dense cell islands in the VS. Together, these data clearly demonstrate that in NHPs, D1-NUDAP and D1/D2-hybrid MSNs are discrete subtypes. However, the limitations involved with integrating single cell data sets[346,347], and the vast differences between the studies – differences that include species, age, single cell technology, transgenic status, MSN sampling density, and sequencing depth – preclude us from determining the role of species in determining the degree of discontinuity between cell types.

In contrast to the distinct boundaries between subtypes, we also observed continuous variation in gene expression[288,289,348]. On a large scale, continuos variation in gene expression was

exemplified by the dorso-ventral gradients of *CRYM* and *CNR1* (Figure 19), but there were also axes of continuous variation within and between MSN subtypes. We examined this variation using archetypal analysis[320-322]. Archetypes have biologically interpretable dimensions and concepts, in genes and cell states, respectively. Moreover, projecting archetypes learned in one biological replicate onto another biological replicate requires only a simple matrix operation. This simplicity enabled us to clearly demonstrate the similarity between the biological replicates (Figure S3G). Within most of the MSN subtypes, we observed several archetypes. For example, archetypal analysis of the matrix clusters revealed an archetype that highlighted VS derived nuclei (Figure S3H). Some of the genes correlated with this VS archetype are upregulted in the NAc core[323]. Thus, this archetype analysis defined a potential NAc core signal. One challenge that remains is to determine whether archetypes indicate subtypes or 'states,' with the former being a stable feature found across individuals and the latter being a transitory phase that could be activity dependent. In the case of the D1Sh *TAC3*-archetype, the fact that we detected it in two monkeys indicates that this archetype is akin to a minor subtype, but other archetypes may indicate transitory cell states. As we gather more data about cell types – for example data on sexual dimorphisms, epigenomic features, physical circuits, and behavioral functions – we believe that the concepts captured by archetypal analysis will be crucial for organizing, describing, and modeling the vast functional heterogenity that charachterizes even simple brain structures like the striatum.

The challenge of defining "cell types" was the genesis of modern neuroscience[349]. More recently, we have come to understand cell types as complex distributions of molecular processes[350,351]. The dynamics of such processes rarely fit simple boundaries, but nonetheless we continue to use the idea of cell types to abstract molecular, neurophysiological, and morphological patterns that we observe in our cells of study. We foresee at least three critical reasons to avidly

continue doing so in NHP. First, Old-world monkeys, like Rhesus macaques, are more similar to humans than any other research animal that allows for invasive neurophysiological experiments. Accordingly, Rhesus macaque cell types, including highly specialized neurons like Betz cells, von Economo neurons, and even striatal interneuron types, recapitulate homologous human cell types better than cells from rodents or even from marmosets[291,352,353]. Second, single cell studies performed on post-mortem human tissue are subject to different ethical constraints that manifest as relatively long and highly variable postmortem intervals[300]. In contrast, NHP experiments can be performed in a highly controlled and more timely fashion. Finally, NHPs have resisted the widespread application of modern genetically coded tools. Single cell technologies, including snRNA-Seq and snATAC-Seq can identify cell types and potent regulatory sequences that will break this resistance and enable the effective applications of genetically coded tools in large, wild type animals that resemble humans. Rhesus macaque behavior is readily interpretable in terms of human behavioral theory such as economic theory[36,37,269,354], learning theory[217], and even game theory[88,355]. The diversity of MSN cell types presented here provides a blueprint to investigate the cell-type-specific mechanisms for such sophisticated behaviors.

5.0 Discussion

Uncertainty is a ubiquitous fact of life that persists in small decisions like choosing which coffee to order in the morning, to incredibly consequential decisions like choosing which house to buy or what job to accept. In order to account for uncertainty in our environment, we know that risk is integrated into the way decision-makers assess the values of choice options, and thus, into brain regions that code estimates of subjective value. Ambiguity, on the other hand, is a less understood form of uncertainty in terms of the impact it has on subjective value coding and learning, and the neural regions that mediate this. However, real world decisions are more reflective of ambiguity than of risk, but most research utilizes risk because of the ability to precisely control a number of variables that are difficult to control when trying to elicit ambiguity. The work discussed here has filled a critical gap in our understanding of how uncertainty effects behavior, how neurons encode complex environments through representations of higher order statistics, and possible future striatal targets for cell type- or circuit-specific manipulations in nonhuman primates (NHPs).

5.1 **Reward Distribution Coding in Midbrain Dopamine Neurons**

Midbrain dopamine neuron reward prediction error (RPE) responses showed enhanced coding to rare rewards compared to the same rewards that were more commonly experienced. The work described here is the first study to show that distributions over rewards can be represented in individual dopamine neuron firing rates with matching reward outcomes, expected values, and prediction errors – as calculated by a standard temporal difference (TD) learning rule where RPE

is equal to the reward received minus the reward predicted. In addition, we corroborated previous studies, showing that distributions over rewards with higher uncertainty caused slower learning compared to distributions with more certainty[122]. While the recordings we performed were in a passive viewing task, and not the choice task used to estimate learning, it is an intuitive jump to suggest that decreases in dopamine responses to common rewards compared to rare rewards of the same magnitude, would elicit smaller dopamine neurons coding in a learning task. Further, reward prediction error signals can be seen in dopamine concentrations in the striatum. Specifically, dopamine release in the nucleus accumbens (NAc) shows graded, bi-directional concentration changes reflective of RPE magnitude, such that larger positive RPEs increase dopamine concentration compared to smaller RPEs, and larger negative RPEs decrease dopamine concentration compared to smaller negative RPEs[356]. This fine-tuned coding of distributions over rewards in midbrain dopamine neurons could indeed also be represented via dopamine release in downstream targets that are critical for value learning and updating behavior.

In addition to the slower learning from distributions over reward with higher uncertainty, phasic changes in pupil diameter were also smaller in comparison to distributions over reward that had higher certainty. Pupil diameter is a physiological measure of norepinephrine release, stemming from the locus coeruleus (LC), which is a sign of sympathetic nervous system activation. It has also been shown that, post-reward feedback, pupil dilation drives learning in uncertain perceptual environments. Midbrain dopamine neurons and the LC have direct and indirect reciprocal connections, and both receive prefrontal cortical inputs[357-359]. Further, LC and midbrain fMRI responses are functionally coupled[360]. While these observations are only correlational, they are certainly grounds for experiments to disentangle the relationship between midbrain dopamine neuron and LC firing rates in response to learning from varying degrees of uncertainty in reward

distributions, sympathetic activation as measured by pupil response, and learning. A working hypothesis is that higher levels of uncertainty in choices causes enhanced bidirectional decreases in the magnitude of positive and negative RPE responses in dopamine neurons, minimized responses in LC neural coding and pupil dilation, and slower learning in choice situations.

A question that should be asked following our observations in Chapter 2 – that rare rewards cause amplified dopamine reward prediction error responses is: "Why do we need a representation of reward uncertainty, and how is it used in downstream brain regions?" Dopamine projects widely through the brain. Areas of particular interest include the striatum and amygdala[361,362], due to their role in reinforcement learning and observable effects following dopamine manipulations. Standard reinforcement learning (RL) models include a representation of the state of the environment as estimated by the expected value of that state after integrating input from the critic. However, this model lacks higher order statistics to describe the state. Partially observable Markov decision process (POMDP) models of RL furthers the standard RL model, such that it operates under the assumption that the agent can create a distribution of the possible outcomes given the state. Specifically, the amygdala and ventral striatum (VS) have been shown to encode both the immediate expected value (IEV) of exploitation of a given state, and the future expected value (FEV) by exploring novel options as related to a POMDP RL[127]. If the POMDP RL model integrates distributions of outcomes, it is safe to say that there must be some biological input to represent this variable. Our finding, that dopamine neurons encoding distributions over reward, could suggest the representation of a distribution over the possible outcomes utilized in the POMDP RL model could be the distributional representation of outcomes that the amygdala and VS utilize to update their representations of IEV. In fact, another study has shown with the use of a POMDP RL model, dopamine neurons reflected belief state about their probability of being

rewarded based on perceptual ambiguity[363]. In regards to the FEV of exploration, it has been shown that in tasks where the values of two options can be inferred from another and has a set number of trials before value reversal, dopamine neurons utilize this inferred value on the first trials of a new block where the cue and outcome had not yet been experienced[136]. Specifically, on the first trial of the new block, dopamine neurons switch their coding from a negative RPE response to the previously unrewarded cue, to a positive RPE response – indicating their representation of the expected value of the target has changed to be one of not rewarded, to rewarded. The inferred RPE response following the cue then strengthens after actual experience of the reward. Further, it has been shown that dopamine release in the VS is reflective of RPE responses – putatively matching the encoding from dopamine cell bodies[123]. While there is a possibility for distributional coding across populations of dopamine neurons[231], there currently is not enough data to make the jump from individual dopamine neurons coding distributions over rewards, to populations of dopamine neurons coding both inferred FEV distributions over rewards and experienced IEV distributions over rewards. However, this certainly leaves room for investigation into the possibility of midbrain dopamine neurons encoding higher order statistics of distributions of rewarding outcomes across populations for experienced rewards and inferred value, and the utility of this in downstream regions like the amygdala and VS.

In Chapter 4 we characterized multiple different subtypes of MSNs across the macaque striatum, a region with dense inputs from dopamine neurons that is crucial in mediating reward-based learning and decision making. One possibility is that a specific subtype, or subtypes, of these MSNs could have the unique function of integrating signals about distributions of rewards, utilizing that information at the local level, and then propagating it to its own downstream targets. The dorsal striatum (DS) striosome and matrix subspaces are poorly understood in terms of their

function. Generally, the striosome and matrix are thought to participate in limbic and sensorimotor functions, respectively[278-280]. One hypothesis is that striosome MSNs, and not matrix MSNs, could receive and functionally integrate information about distributions of rewards, and that this could be evident in their electrophysiological activity. Utilizing cell type-specific optogenetics in order to photo-tag DS D1 striosome neurons and D1 matrix neurons during a learning task with distributions of rewards (like the learning task in Chapter 2), would be incredibly enlightening in order to compare and contrast the effects of dopamine neuron coding over distributions of rewards on these distinct types of MSNs.

5.2 **Uncertainty Preferences, Ambiguity Aversion and Potential Neural Mediators**

Individual decision-makers have varying attitudes about uncertainty in different contexts. Further, individuals are typically so reluctant to choose ambiguous options, so much so that they would rather choose a gamble where they know the probabilities, even if the expected value is lower than an ambiguous option. The results from Chapter 3 suggest that even when expected value, range of rewards, and amount of uncertainty were matched, there was still a behavioral and sympathetic difference in the effect of ambiguity versus risk. By matching all of these variables, the question of what mechanisms are causing ambiguity aversion still remain, and what manipulation is necessary to parse out the cause of this maladaptive behavior. One possibility is that, under ambiguous choice conditions, decision-makers have a difficult time making predictions about the expected value of the cue, which will lead to noisy value estimates, minimizing their confidence in their choices. While perceptual uncertainty and reward uncertainty are surely arising from different issues of estimation, uncertainty in the perceptual domain or in the computation of

reward estimates, respectively, we should not hesitate in attempting to generalize some of the methods from the vast literature in perceptual decision making. Many studies investigating perceptual decision making have utilized choice confidence into their belief state regarding their probability of choosing correctly, as the decision-makers must discriminate noisy stimuli, like random dot motion stimuli[364], or odors[363]. In addition, decision-makers use these internal belief states about the perceptual uncertainty of an environment to adjust their choice behavior[365-367]. Future studies attempting to parse the neural causes of ambiguity aversion as it relates to estimates of the subjective values of cues and their underlying reward probabilities, could utilize a post-decision wager, to determine the decision-maker's confidence in their estimates of value. In a study utilizing perceptual decision making with measures of choice confidence following decisions, the authors intuit that under perceptual uncertainty, the decision-maker must utilize estimates of belief state, which is their subjective estimate of the true state of the environment based on their current perceptual experience[363]. Further, they showed that midbrain dopamine neurons were sensitive to decision confidence, and integrated their confidence about their perceptual discrimination such that the probability they believed they were correct, and would subsequently be rewarded, influenced midbrain dopamine reward prediction errors. These results combined with ours, suggests that the model utilized in the aforementioned study[363] could be used to estimate belief state in regards to underlying reward distribution, and that this signal could be detectable in midbrain dopamine neuron firing rates.

Ambiguity aversion is highly individual-specific, so much so that it has been shown that optimistic people are less ambiguity averse than pessimistic people, and therefore make better decisions in ambiguous environments[368]. Sub-optimal decision making is a hallmark of a number of psychiatric disorders that also influence affect, such as schizophrenia, anxiety, bipolar disorder,

and depression[369-371]. Critically, dopamine is a key pharmacological target in treating the disorders previously mentioned. Medications used to manage the positive symptoms of schizophrenia – like delusions or hallucinations – target dopamine[372]. Gaining an understanding in dopamine cell body firing – and their subsequent release of dopamine in downstream targets – in response to ambiguity could provide insights into sub-optimal decision making in ambiguous environments and provide understandings into the possible disruptions in disorders characterized by poor decision making. Specifically, in psychiatric or neurodegenerative diseases where dopamine is a therapeutic target.

There has been previous research into the effect of 'advanced information' on decision making preferences and neural activity in midbrain dopamine neurons in non-human primates (NHPs)[373]. Speculation may lead one to believe that the risky cues used here have advanced information about the distribution, and thus the ambiguous cues could present a similar avoidance due to the lacking probability information as seen in the aforementioned study. However, there are three things that could counteract this notion. The first has already been seen, as monkeys in this study never have a preference for the no information cues, whereas in our work, we see that monkeys do indeed prefer ambiguity when the EV is on the lower end of the value bars – or in other words, when the stakes are low (Fig. 14a, left).

Second, it is possible that the advanced information preference seen in the previously mentioned study is a result of a discounting delay in RPEs when monkeys have to wait to experience the rewarding outcome of the trial, instead when they receive the instructional cue that reveals what the reward will be and elicits a RPE response. In other words, it is known that dopamine neurons incorporate delay discounting into RPE coding[374]. In the case of advanced information about the outcome of the trial, they are experiencing the RPE sooner than if they would have had to wait in the no information condition. Dopamine RPE responses occur initially at the

time of an unexpected reward, but as an animal learns that a cue predicts a reward, the dopamine RPE response occurs following the cue and not the reward, because it is no longer unexpected. A smaller delay period to the RPE would create an enhanced dopamine coding, as it is of more subjective value – or utility[375] – due to delay discounting imposed on RPE responses[374,376]. Further, sooner RPE responses would cause release dopamine sooner and in greater amounts to downstream projections. As shown in Chapter 2, dopamine neurons incorporate higher statistical moments into RPE coding that are independent from the expected value and thus, the magnitude of the calculated reward prediction error – which only incorporates the received reward minus the expected reward[377]. It could be inferred from previous research that delay may also be incorporated into reward prediction error coding as an independent aspect of expected value. Specifically, in the paper that defined the effect of delay discounting on dopamine neurons RPE responses, it was revealed that monkeys were choice indifferent to two stimuli with different reward volumes and delay times – a small reward with a short delay, and a large reward with a long delay[374]. The authors showed there were slightly smaller overall dopamine responses to the cue with a longer delay, even though they had the same expected value. This suggests a specific delay-induced minimization of dopamine reward prediction error responses, or possibly baseline activity, that is separate from expected value. It has been shown that dopamine responses and subsequent release are necessary and sufficient for effective learning and updating of value[43,378-382], which is theorized to be through Hebbian learning mechanisms[383]. Instituting delays to rewards following predictive cues also has been shown to reduce the speed with which animals can learn these dopamine-mediated cue-reward associations[376]. Thus, the reduction in dopamine firing, and subsequent release, following a delayed RPE in the 'No Information' condition could be the source of reduced valuation of that cue compared to the faster RPE experienced in the 'Information' condition.

Further, in terms of reinforcement learning, the delay to knowing the outcome in the no information condition would delay credit assignment to the stimulus like the mPFC, which receives dopamine innervation, to continue firing to preserve the memory of the value of the action the animal performed to receive the reward, perhaps weakening the representation of the 'No Information' cue value in later trials[69,80]. These are all possibilities for why the animals would hold a preference for the faster RPE experienced in the advanced information condition[373].

Finally, a behavioral experiment that could determine if this was indeed an effect of advanced information received from the risky cues, as opposed to a genuine effect of ambiguity, would be to include a fractal cue in the choice task seen in Chapter 3, that also predicts a uniform distribution – thus having the same expected value, reward distribution, and uncertainty as the ambiguous cue, and the Uniform value bar cue. This finding would not only disprove the idea that the ambiguity the monkeys experience here is a difference of advanced information, but instead, a psychological state induced by our manipulation by 'hiding' information about reward probabilities that they typically can access in the risky cues.

It has been shown that learning about the Ellsberg Paradox reduces ambiguity aversion, thus increasing behavioral performance in humans[255]. This finding shows that while this is an engrained bias, it can be overcome using cognitive inhibition. This leaves room to investigate what brain areas perform top-down modulations to decision making in economic decisions with ambiguity. Some areas that have previously been shown to be differentially responsive to ambiguity versus risk in economic decision making are the lateral orbitofrontal cortex (lOFC) and the amygdala. A previous study using fMRI in humans performing a decision making task show increases in lOFC and the amygdala in ambiguous decisions compared to risky ones[274]. Further,

the striatum was more activated in risky decisions over ambiguous ones. These results provide the basis for further behavioral and neuronal characterizations of decision making under ambiguity.

Due to the unique subjective experience of risk and ambiguity that elicits very different – and sometimes illogical – behavioral patterns across individuals, perhaps it is possible that while DS D1 striosome MSNs mediate information about distributions over rewards, a different subtype of MSNs could mediate our preferences in uncertain environments. Specifically, there are "hedonic hotspots" located in the nucleus accumbens and ventral pallidum that, when opioids are directly applied, produce behavioral reactions that indicate pleasure[343]. One possibility is that the MSNs we identified as NUDAPs could be a part of these hedonic hotspots, and these areas mediate our uncertainty preferences. Uncertainty – or risk and ambiguity – can be pleasurable, even to a pathological degree, as can be seen by compulsive gambling, and other addiction disorders. But uncertainty can also induce negative feelings and distress, again, which can become pathological in disorders like anxiety or depression. The striatum is such a critical region for learning, and hedonic hotspots could mediate the subjective interpretation of objective uncertainty that produces individual-specific preferences.

5.3 **Final Thoughts & Future Directions**

The most important result to utilize moving forward is that a standard Temporal Difference (TD) reward prediction error (RPE) model is not sufficient to account for the RPE responses seen in dopamine neurons. Our results

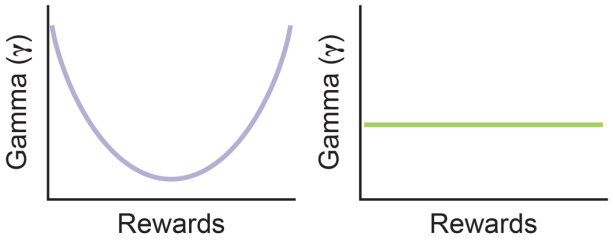


Figure 24: Uncertainty term, gamma (γ), for Normal and Uniform distributions

illustrate that second order statistics, and possibly the full probability distribution over rewards, are incorporated into dopamine RPE responses. There are RL algorithms that may be more adept at describing the results we have shown, such as the Kalman TD[232]. We propose a framework similar to Kalman TD, where the estimation of uncertainty (γ) would be the inverse of the probability of receiving that reward – such that a normal distribution would have an uncertainty term that represents an inverted U-shape over possible rewards, and a uniform distribution would have a constant term as the outcome are all equally probable (Fig. 24). The reward prediction error on every trial would then be calculated by taking the product of a standard TD RPE and the uncertainty term generated based on underlying probability distribution over rewards, and would update the value of the chosen option as follows:

$$V_{t+1} = V_t + \alpha * (\delta_t * \gamma)$$

(Eq. 9)

where V is the value of the chosen option, t is the current trial, and α is the learning rate. The Kalman TD is the product of δ_t, which is a TD RPE, and γ, the term representing the inverse of the likelihood of receiving the reward that was given (Fig. 24). An important note is that we do not attribute the amplification we saw in dopamine neurons to be coding the product of the standard TD RPE and the learning rate term. This is because in our recordings the CS-reward associations were stable over many thousands of trials, and thus the effective learning rate was likely near zero. Therefore, we postulate that there must be a specific term for the uncertainty of an environment, or γ, separate from the learning rate term, α. The observed amplification of dopamine responses by rare rewards is consistent with a signal that could guide Bayesian inference of the most likely outcomes. While our results indicate that this algorithm is a strong possibility in a learned, yet risky, environment, further research is required to understand how these dopamine signals change

during the context of learning, and what our baseline estimation of the underlying distribution over rewards is. In other words, how do we learn the second order statistics about distributions of rewards following a choice? Future research will be required to determine whether a Kalman TD is an accurate representation of dopamine reward prediction error signals during learning in risky environments with underlying probability distributions over multiple rewards. Further, discovering the assumptions made by decision-makers about the probabilities of rewards before learning the true underlying distribution will be important in understanding how we learn statistics of rewarding environments. There are multiple instances of neural coding taking the shape of a normal distribution in the brain, such as orientation tuning in V1, that have a normal distribution coding scheme with the peak at the neurons' preferred orientation[384]. If this is the standard model used in neural coding and the assumption made about the statistics of our environment, this could be a contributing factor as to why learning from narrower, normal distributions is easier than broader, uniform distributions – as it is already an innate feature of our brains.

While each of these results are individually interesting and provide a novel contribution to neuroscience, there is a common through line that ties them all together. Dopamine neurons send learning signals to their downstream projections, such that phasic activations teach us to repeat behaviors, and phasic suppression teach us to stop behaviors. Dopamine neurons are less sensitive to common RPE responses from uncertain distributions, and learning is slowed by uncertainty. Additionally, we saw that broader, more uncertain distributions slowed learning, and that ambiguity was treated as 'super risk', such that ambiguity preferences in low and high stakes choices matched risk preferences, but were much more exaggerated in comparison to risky cues with the same uncertainty. We saw that animals had larger pupil responses in more certain normal distributions during learning, and they also learned faster. In a choice context with no learning

involved, pupil responses were smaller in ambiguous compared to risky decisions and outcomes. Based on these observations, ambiguity aversion could be caused by dopamine neurons. If ambiguous cues and rewards elicit a smaller magnitude of RPE responses in dopamine neurons, this could slow the value learning needed to accurately choose the best option, which would also be reflected in minimized pupil diameter responses. In low stakes contexts, this explain why we see such a preference for ambiguity, because like with risk preferences, they are not concerned with losing smaller volumes of safe juice, and indeed value the uncertainty surrounding the ambiguous or risk cue – even though they may not have an accurate estimation of the underlying value or distribution over rewards. However, in the high stakes conditions, decision-makers are not likely to gamble with larger reward volumes and, indeed, value a higher value of certainty, as we saw with the exaggerated undervaluing of the ambiguous cue. Furthermore, now that we have characterized unique MSNs in the primate striatum – which all receive dense projections from dopamine neurons, it is indeed plausible that a specific subtype of these MSNs might integrate information about distributions of rewards into their coding. Specifically, D1 striosome neurons of the striatum could incorporate information about the distributions over rewards in a particular environment, while NUDAPs that possibly make up hedonic hotspots color our individual experience and preferences towards the uncertainty of these distributions over rewards.

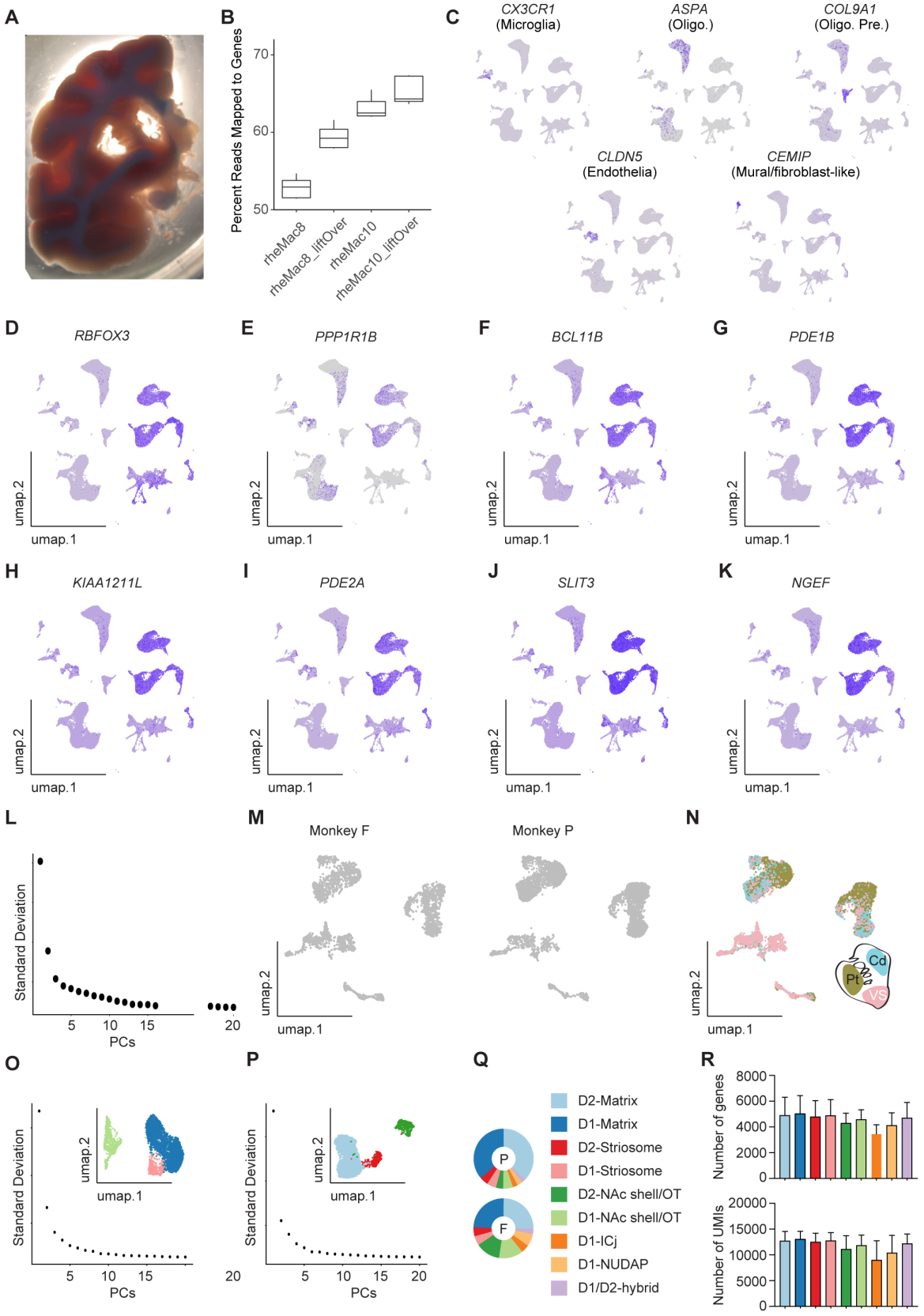
Together, this body of work has provided information that compels us to update a long-held standard for how we think of the coding of dopamine neurons, and clear directions for future research. Key questions include understanding how the brain uses information about the second order statistics of rewarding environments encoded by dopamine neurons, and how this relates to ambiguity aversion. Further, newly characterized medium spiny neuron types in the macaque

striatum could be possible points of investigation in terms of how we incorporate information about distributions over rewards, and our subjective preferences for uncertainty.

Appendix A Supplemental Information

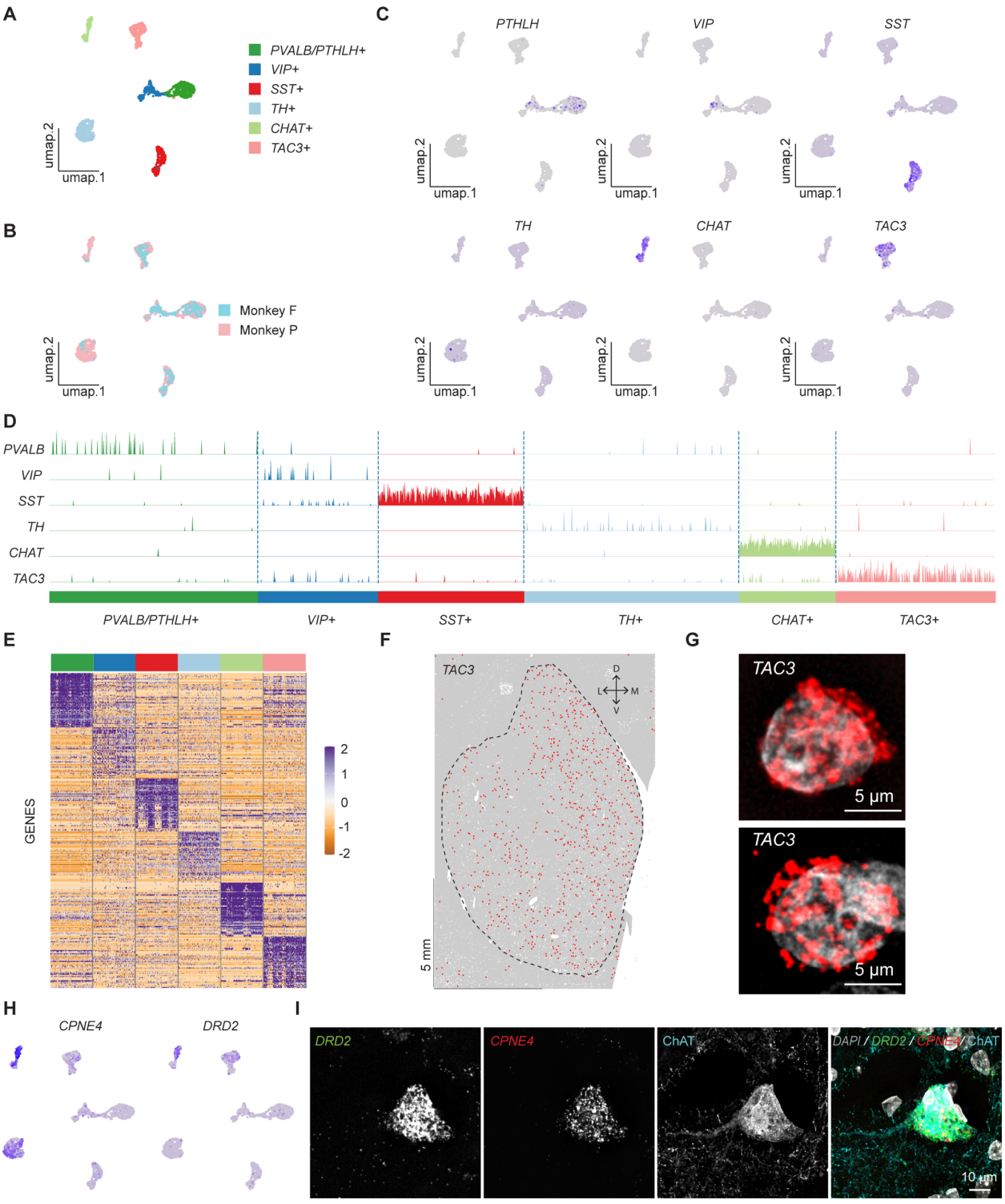
Supplemental data and tables for Chapter 4 can be found online at https://doi.org/10.1016/j.cub.2021.10.015.

Supplemental Figures



Supplementary Figure 1: Read Mapping, Feature Plots, Sub-clustering, and Gene Expression Counting, Related to Figure 17 and 18

A, A slab of monkey tissue after caudate and putamen were dissected out.
B, Comparison of gene mapping rate using four different gtf in our alignment of reads to macaque genome. We used UCSC liftOver tool to liftOver the human transcriptome gtf file onto the rheMac8 or rheMac10 genome to get liftOver version rheMac8_liftOver and rheMac10_liftOver, respectively.
C, Feature plots of marker gene expression for major cell classes in striatal nuclei.
D, Feature plot showing the expression of *RBFOX3*, which labels all neurons.
E-K, Feature plots of gene expression for well-known MSN markers, *PPP1R1B* and *BCL11B*, *PDE1B* and new MSN-specific markers, *KIAA1211L*, *PDE2A*, *SLIT3*, *NGEF*.
L, Standard deviation in different principal components (PCs) for all MSNs. The first 15 PCs account for majority of the variation.
M, Side by side monkey comparison for MSNs in UMAP coordinates.
N, UMAP visualization of MSNs colored by three regions.
O, Standard deviation in different principal components (PCs) for D1 MSNs. Inset showing the same spatially distinct UMAP clusters (D1-Matrix, D1-Striosome, and D1-Shell/OT) were recovered after isolating *DRD1*-MSNs and then performing PCA and UMAP dimensionality reduction. Color scheme is the same as in **Q**.
P, Standard deviation in different principal components (PCs) for D2 MSNs. Inset showing the same spatially distinct UMAP clusters (D2-Matrix, D2-Striosome, and D2-Shell/OT) were recovered after isolating *DRD2*-MSNs and then performing PCA and UMAP dimensionality reduction. Color scheme is the same as in **Q**.
Q, Distribution of relative proportion of each MSN type in monkey P and F.
R, Number of genes (top) and unique molecular identifiers (UMIs, bottom) across nine MSN types. Error bars indicate standard deviation across two monkeys.



Supplementary Figure 2: Interneurons in Primate Striatum, Related to Figure 17

A, Re-clustering of interneurons revealed six distinct interneuron clusters that corresponded to known populations of striatal interneurons (*PTHLH*/*PVALB*-, *VIP*-, *SST*-, *TH*-, *CHAT*-, and *TAC3*-positive interneurons).

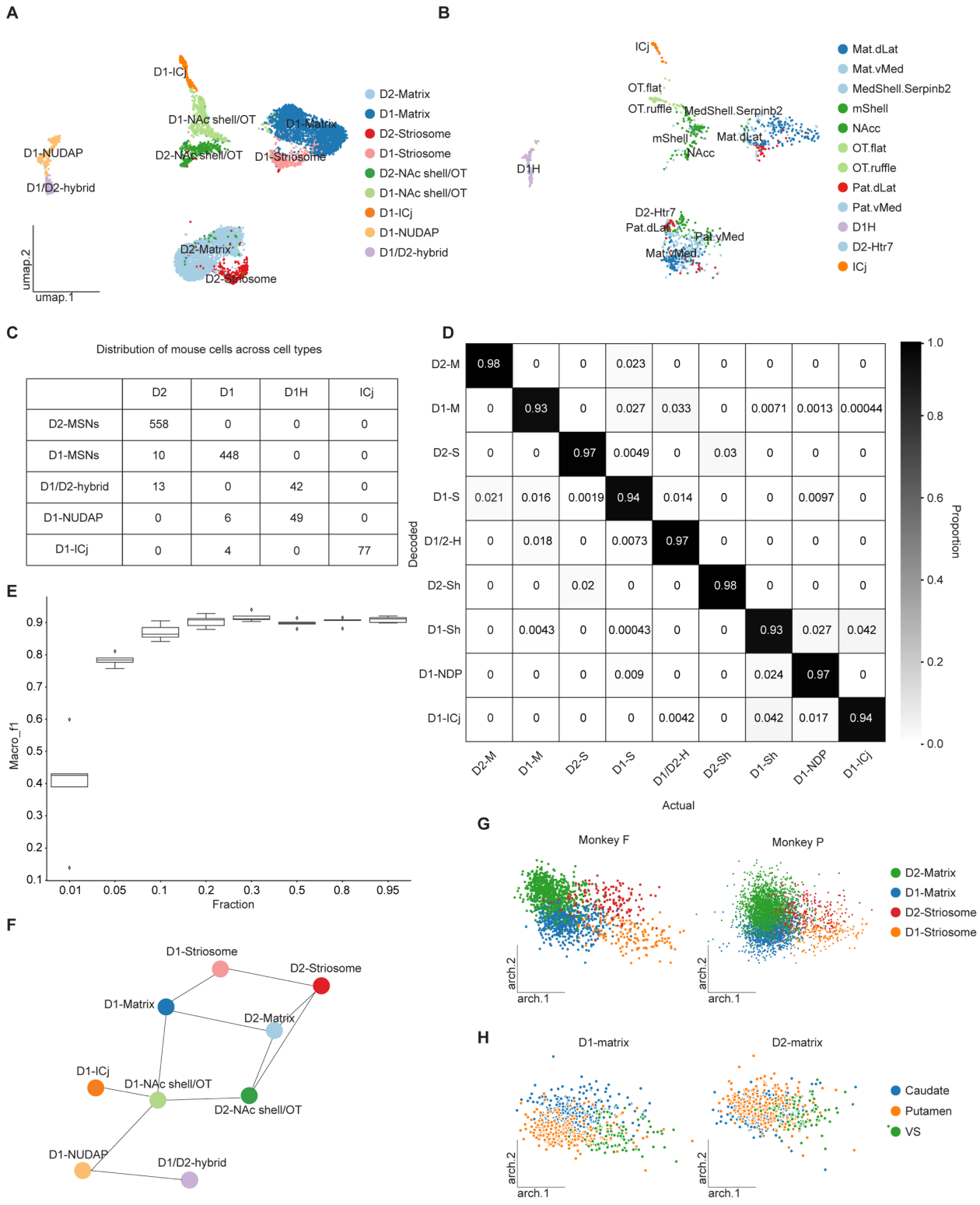
B, UMAP visualizations of the samples from the two subjects.

C, Feature plots of interneuron marker genes for each subtype.

D, TrackPlot of marker genes in each subtype.

E, Heatmap of differentially expressed genes across interneuron subtypes.

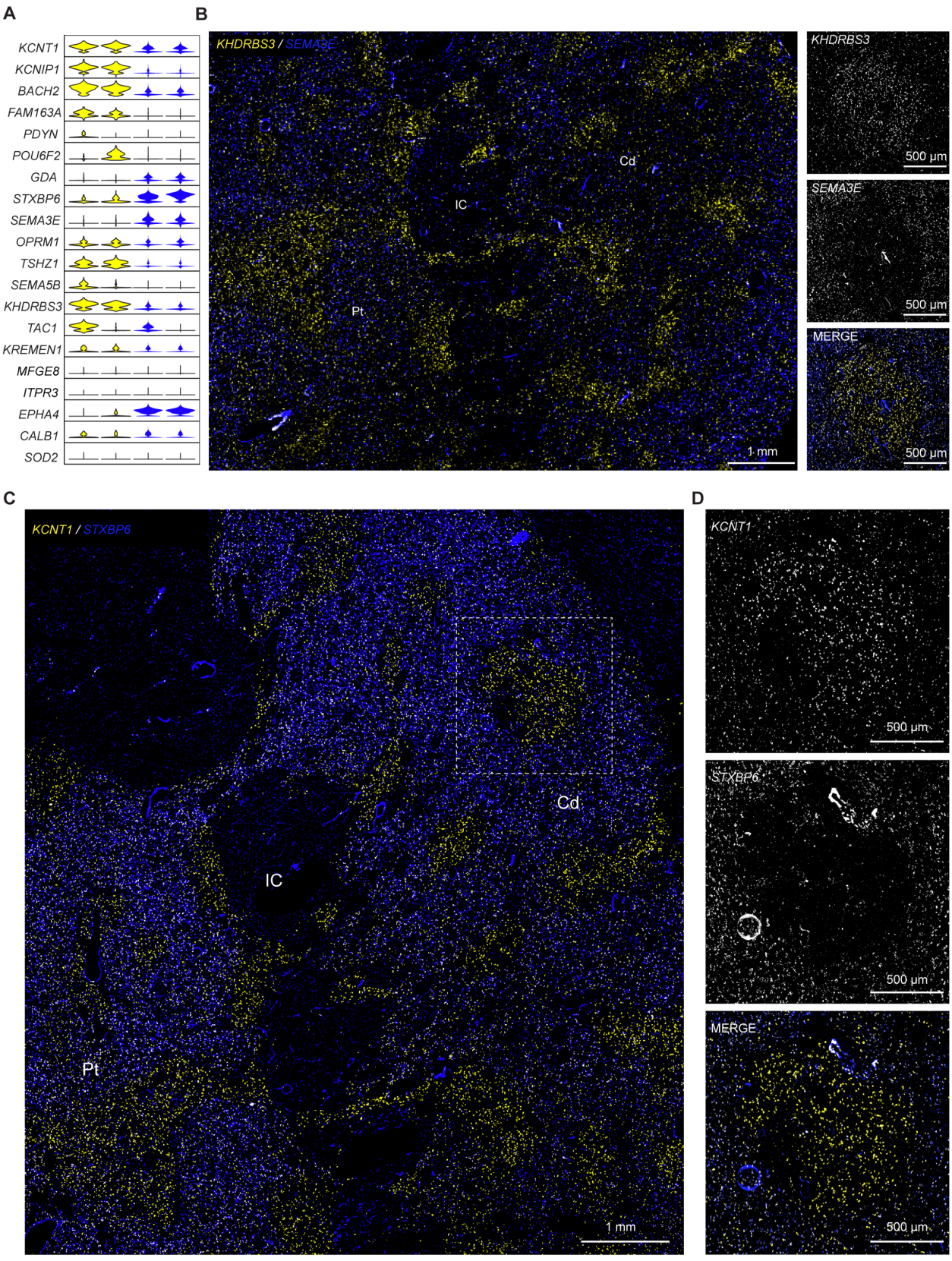
F, FISH labeling for *TAC3* – a marker gene for a novel class of interneurons that are unique to primates[291] – indicated that these interneurons are broadly but sparsely distributed throughout the Rhesus macaque striatum.
G, Two example *TAC3*-positive cells.
H, The snRNA-Seq data indicated that *CPNE4* and *DRD2* were significant marker genes for cholinergic interneurons.
I, *CPNE4* co-localizes with *DRD2* and ChAT antibody labeling in cholinergic interneurons.



Supplementary Figure 3: Monkey and Mouse MSN Comparison, Confusion Matrix Among MSN types and Archetype analysis, Related to Figures 18 and 19

A, UMAP visualization of monkey subtypes after integration of monkey and mouse MSNs in UMAP coordinates.
B, UMAP visualization of mouse subtypes after integration of monkey and mouse MSNs in UMAP coordinates.

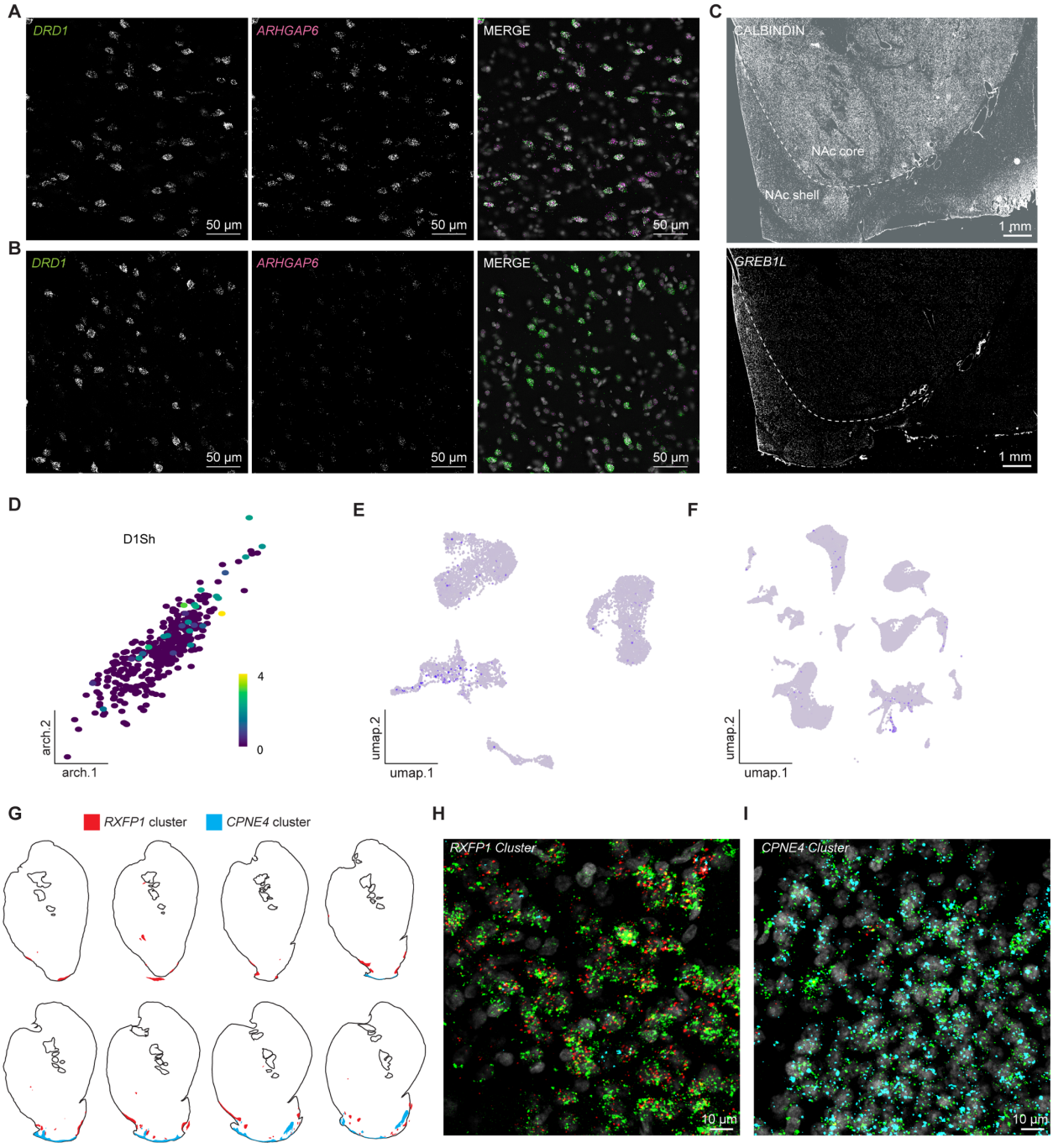
C, Combined monkey and mouse MSNs were re-annotated to (D2-MSNs, D1-MSNs, D1/D2 hybrid, D1- NUDAP and D1-ICj) and the number of cells per mouse MSN cell type that fell within the re-annotated clusters was quantified. To make a comparable comparison, mouse cell types were annotated with D1, D2, D1H and ICj as in the paper Stanley et al.[288]
D, The accuracy rate (numbers within the grid) between SCCAF decoded cell type and actual cell type using the data combined from both subjects.
E, f1 score from SCCAF after down sampling. Systematic down-sampling of the MSN samples revealed that we had adequately sampled the underlying heterogeneity: even dividing the sample in half, or more, had little impact on our ability to decode the subclusters.
F, MSN subtypes connected by PAGA analysis
G, The calculated archetypes weights – trained on the data from one monkey showed the same structure for the subtypes in both subjects.
H, Archetype analysis on the D1- and D2-matrix showed an archetype originated from ventral striatum (VS).



Supplementary Figure 4: Striosome and Matrix Labeled by FISH Markers, Related to Figure 20

A, Violin plot showing our identified and published striosome and matrix markers. From left to right columns are D1

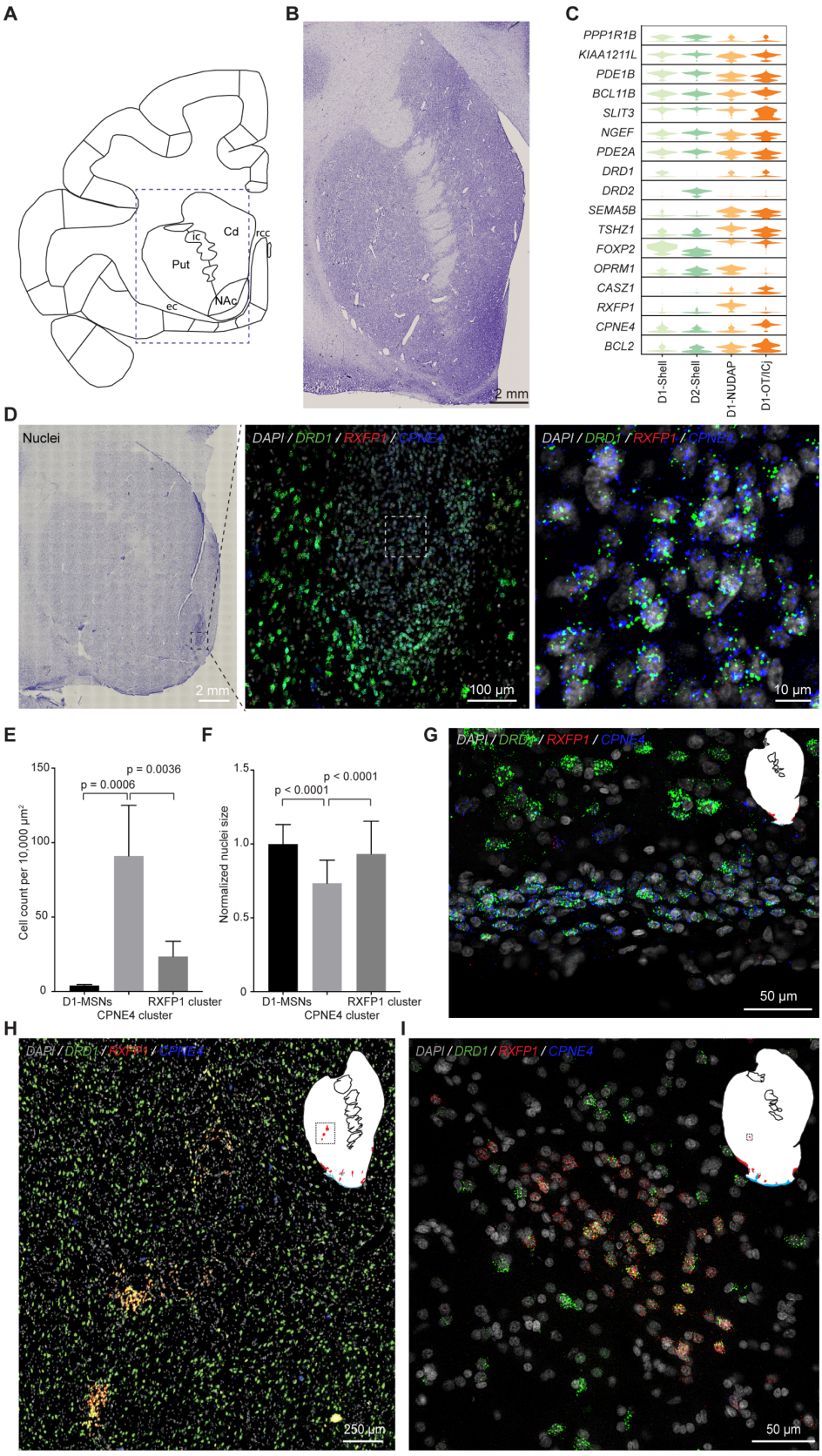
striosome, D2 striosome, D1 matrix and D2 matrix. *PDYN* is a specific D1 striosome marker and *POU6F2* is a specific D2 striosome marker.
B, FISH labeling of *KHDRBS3* (yellow) and *SEMA3E* (blue) showing clear striosome and matrix distinction. Right images show representative striosome and matrix from left image.
C, FISH labeling of *KCNT1* (yellow) and *STXBP6* (blue) showing striosome and matrix compartmentation. Cd: caudate, IC: internal capsule, Pt: putamen.
D, Detail of the white square in C. Scale bars are indicated on the images.



Supplementary Figure 5: *RXFP1* and *CPNE4* Cluster Distribution in a Second Monkey, Related to Figures 21 and 22

A, High-resolution image of the ROI indicated with the letter "A" in Figure 21A.
B, High-resolution image of the ROI indicated with the letter "B" in Figure 21A.
C, (top) Calbindin immunohistochemistry reveals the border between the core and shell (dashed white line). (bottom) FISH labelling of *GREB1L* and *GREB1L* intensity follows the border of the shell (dashed white line).
D, *TAC3* expressions along the D1 shell/OT archetype axes.
E, UMAP plot of *TAC3* in MSNs.
F, UMAP plot of *TAC3* in nuclei including all major cell classes.
G, Distribution of *RXFP1* and *CPNE4* clusters across eight rostral-caudal regions identified by multichannel FISH in a second monkey K.

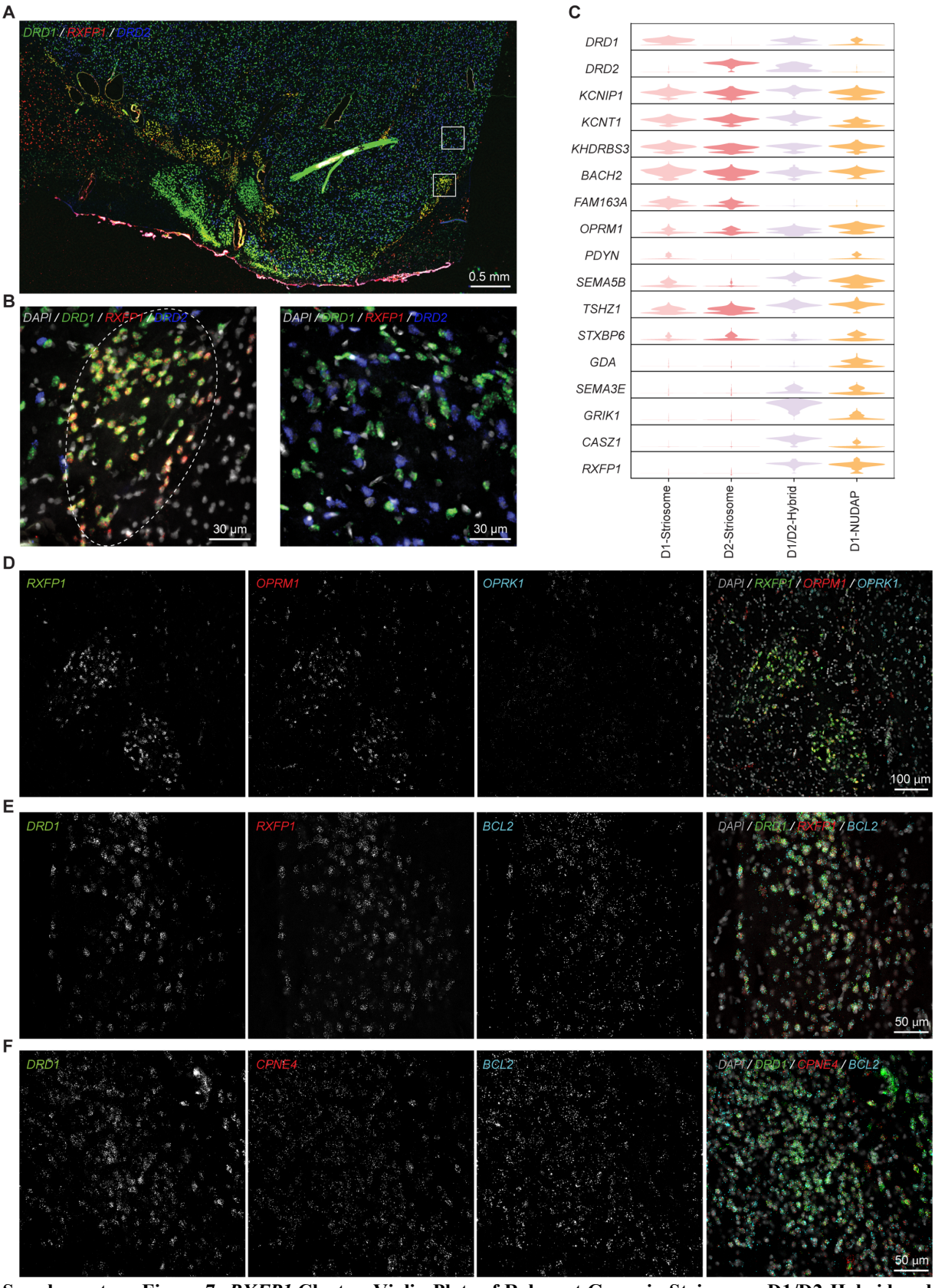
H, FISH stain of *DRD1* (green), *RXFP1* (red) and *CPNE4* (cyan). Nuclei were labeled by DAPI (grey). This image shows a representative *RXFP1* cluster in the ventral striatum.
I**,** FISH stain of *DRD1* (green), *RXFP1* (red) and *CPNE4* (cyan). Nuclei were labeled by DAPI (grey). This image shows a representative *CPNE4* cluster in the ventral striatum.



Supplementary Figure 6: Comparing Nissl Section with Paxinos Atlas and *RXFP1* and *CPNE4* Cluster Mapping, Related to Figure 22

A, Paxinos atlas shows the ventral extent of the external and extreme capsules loop around the ventral portion of the

NAc and connects to the rostrum of the corpus callosum. Cd: caudate, Put: putamen, ec: external capsule, NAc: nucleus accumbens, rcc: rostrum of the corpus callosum, ic: internal capsule.
B, Nissl stain of one section. This Nissl image corresponds to the blue dashed box on the left.
C, Violin plots of relevant marker genes in ventral striatum MSN types.
D, Nissl stain of a section (left) and triple labeling of *DRD1* (green), *RXFP1* (red) and *CPNE4* (blue) in an adjacent section (middle and right) shows that *CPNE4* labels major island of Calleja. Nuclei labeled by DAPI (grey). Right image is the enlarged image from the boxed region in the middle image.
E, Quantification of cell density of neurons expressing *DRD1* in *RXFP1* and *CPNE4* clusters and nearby MSNs. One-way ANOVA with Bonferroni post hoc test was used for statistical analysis. Error bars are SD across 4 sections.
F, Quantification of nuclei size of neurons expressing *DRD1* in *RXFP1* and *CPNE4* clusters and nearby D1- MSNs. Nuclei size was normalized to the mean area size of regular D1-MSNs in each section. One-way ANOVA with Bonferroni post hoc test was used for statistical analysis. Error bars are SD across 49 cells from three sections.
G, FISH labeling of *DRD1* (green), *RXFP1* (red) and *CPNE4* (blue) in a section. Nuclei labeled by DAPI (grey). Inset: white area indicates striatum and the dashed black box highlights the area shown in the FISH image. The bottom *CPNE4* cluster mapped to AON.
H, FISH stain of *DRD1* (green), *RXFP1* (red), and *CPNE4* (blue) in one of section from monkey B. Inset: white area indicates striatum and the dashed black box highlights the area shown in the FISH image.
I, FISH stain of *DRD1* (green), *RXFP1* (red), and *CPNE4* (blue) in one of section from monkey K. Inset: white area indicates striatum and the dashed black box highlights the area shown in the FISH image.



Supplementary Figure 7: *RXFP1* Cluster, Violin Plots of Relevant Genes in Striosome, D1/D2-Hybrid and D1-NUDAP Cells and *OPRK1* and *BCL2* expression in NUDAPs, Related to Figures 22 and 23

A, Triple FISH labeling of *DRD1* (green), *RXFP1* (red) and *DRD2* (blue) shows that *RXFP1* islands do not express

DRD2. Black asterisk indicates staining artifact probably from dust. The artifact could be easily differentiated from real signals because there was no individual grain inside the artifact.
B, High resolution images from the boxed regions in **A**.
C, Violin plots of relevant marker genes in D1-striosome, D2-striosome, D1/D2-hybrid, and D1-NUDAP cells.
D, Triple FISH labeling of *RXFP1* (green), *OPRM1* (red) and *OPRK1* (cyan) shows reduced *OPRK1* expression in *RXFP1* islands.
E, Triple FISH labeling of *DRD1* (green), *RXFP1* (red) and *BCL2* (cyan) shows that *RXFP1* islands express neuronal immature marker *BCL2*.
F, Triple FISH labeling of *DRD1* (green), *CPNE4* (red) and *BCL2* (cyan) shows that *CPNE4* islands express neuronal immature marker *BCL2*.

Bibliography

1 Ellsberg, D. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics* **75**, 643-669 (1961).

2 Rumsfeld, D. (Statements of Donald H. Rumsfeld, Secretary of Defense, during a DoD News Briefing on February 12, 2002, Washington, D.C., 2002).

3 Bernoulli, D. Exposition of a New Theory on the Measurement of Risk. **22**, 23-36, doi:10.2307/1909829 (1954).

4 Knight, F. H. *Risk, uncertainty and profit*. (Houghton Mifflin, 1921).

5 Von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior*. (Princeton University Press, 1944).

6 Savage, L. J. *The foundations of statistics*. (Courier Corporation, 1972).

7 Kahneman, D. & Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**, 263-291, doi:10.2307/1914185 (1979).

8 Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. (MIT press, 2018).

9 Thorndike, E. L. *Animal Intelligence: Experimental Studies*. (Macmillan Press, 1911).

10 Pavlov, I. P. Conditioned Reflexes: oxford University Press. *London, UK [Google Scholar]* (1927).

11 Minsky, M. L. *Theory of neural-analog reinforcement systems and its application to the brain-model problem*. (Princeton University, 1954).

12 Farley, B. & Clark, W. d. Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory* **4**, 76-84 (1954).

13 Andreae, J. H. STELLA: A scheme for a learning machine. *IFAC Proceedings Volumes* **1**, 497-502 (1963).

14 Andreae, J. H. & Cashin, P. M. A learning machine with monologue. *International Journal of Man-Machine Studies* **1**, 1-20 (1969).

15 Michie, D. Trial and error. *Science Survey, Part* **2**, 129-145 (1961).

16 Michie, D. Experiments on the mechanization of game-learning Part I. Characterization of the model and its parameters. *The Computer Journal* **6**, 232-236 (1963).

17 Jh, H. Adaptation in natural and artificial systems. *Ann Arbor* (1975).

18 Holland, J. H. Escaping brittleness: the possibilities of general purpose learning algorithms applied to parallel rule-based system. *Machine learning*, 593-623 (1986).

19 Fel'dbaum, A. A. *Optimal Control Systems by AA Fel'Dbaum*. (Elsevier, 1966).

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

21 Bellman, R. Dynamic programming. *Princeton, USA: Princeton University Press* **1**, 3 (1957).

22 Bellman, R. A markov decision process. journal of Mathematical Mechanics. (1957).

23 Howard, R. A. Dynamic programming and markov processes. (1960).

24 Samuel, A. (McGraw-Hill, 1959).

25 Klopf, A. H. *Brain function and adaptive systems: a heterostatic theory*. (Air Force Cambridge Research Laboratories, Air Force Systems Command, United …, 1972).

26 Michie, D. & Chambers, R. A. BOXES: An experiment in adaptive control. *Machine intelligence* **2**, 137-152 (1968).
27 Watkins, C. J. C. H. Learning from delayed rewards. (1989).
28 Watkins, C. J. & Dayan, P. Q-learning. *Machine learning* **8**, 279-292 (1992).
29 Louie, K., Glimcher, P. W. & Webb, R. Adaptive neural coding: from biological to behavioral decision-making. *Curr Opin Behav Sci* **5**, 91-99, doi:10.1016/j.cobeha.2015.08.008 (2015).
30 Ehringer, H. & Hornykiewicz, O. Verteilung von Noradrenalin und Dopamin (3-Hydroxytyramin) im Gehirn des Menschen und ihr Verhalten bei Erkrankungen des extrapyramidalen Systems. *Klinische Wochenschrift* **38**, 1236-1239 (1960).
31 Hassler, R. The pathology of paralysis agitans and post-encephalitic Parkinson's. *Journal fur Psychologie und Neurologie* **48**, 387-476 (1938).
32 Poirier, L. J. Experimental and histological study of midbrain dyskinesias. *Journal of Neurophysiology* **23**, 534-551 (1960).
33 Schultz, W. Depletion of dopamine in the striatum as an experimental model of Parkinsonism: direct effects and adaptive mechanisms. *Prog Neurobiol* **18**, 121-166, doi:10.1016/0301-0082(82)90015-6 (1982).
34 Schultz, W., Studer, A., Jonsson, G., Sundstrom, E. & Mefford, I. Deficits in behavioral initiation and execution processes in monkeys with 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-induced parkinsonism. *Neurosci Lett* **59**, 225-232, doi:10.1016/0304-3940(85)90204-6 (1985).
35 Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593-1599, doi:10.1126/science.275.5306.1593 (1997).
36 Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect marginal utility. *Curr Biol* **24**, 2491-2500, doi:10.1016/j.cub.2014.08.064 (2014).
37 Lak, A., Stauffer, W. R. & Schultz, W. Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proc Natl Acad Sci U S A* **111**, 2343-2348, doi:10.1073/pnas.1321596111 (2014).
38 Kobayashi, S. & Schultz, W. Influence of reward delays on responses of dopamine neurons. *J Neurosci* **28**, 7837-7846, doi:10.1523/JNEUROSCI.1600-08.2008 (2008).
39 Pasquereau, B. & Turner, R. S. Limited encoding of effort by dopamine neurons in a cost-benefit trade-off task. *J Neurosci* **33**, 8288-8300, doi:10.1523/JNEUROSCI.4619-12.2013 (2013).
40 Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898-1902, doi:10.1126/science.1077349 (2003).
41 Bayer, H. M., Lau, B. & Glimcher, P. W. Statistics of midbrain dopamine neuron spike trains in the awake primate. *J Neurophysiol* **98**, 1428-1439, doi:10.1152/jn.01140.2006 (2007).
42 Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive coding of reward value by dopamine neurons. *Science* **307**, 1642-1645, doi:10.1126/science.1105370 (2005).
43 Stauffer, W. R. *et al.* Dopamine Neuron-Specific Optogenetic Stimulation in Rhesus Macaques. *Cell* **166**, 1564-1571 e1566, doi:10.1016/j.cell.2016.08.024 (2016).
44 Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* **16**, 966-973, doi:10.1038/nn.3413 (2013).
45 Chang, C. Y. *et al.* Brief optogenetic inhibition of dopamine neurons mimics endogenous

negative reward prediction errors. *Nat Neurosci* **19**, 111-116, doi:10.1038/nn.4191 (2016).
46 Chang, C. Y., Gardner, M. P. H., Conroy, J. C., Whitaker, L. R. & Schoenbaum, G. Brief, But Not Prolonged, Pauses in the Firing of Midbrain Dopamine Neurons Are Sufficient to Produce a Conditioned Inhibitor. *J Neurosci* **38**, 8822-8830, doi:10.1523/JNEUROSCI.0144-18.2018 (2018).
47 Araki, M., McGeer, P. L. & Kimura, H. The efferent projections of the rat lateral habenular nucleus revealed by the PHA-L anterograde tracing method. *Brain Res* **441**, 319-330, doi:10.1016/0006-8993(88)91410-2 (1988).
48 Herkenham, M. & Nauta, W. J. Efferent connections of the habenular nuclei in the rat. *J Comp Neurol* **187**, 19-47, doi:10.1002/cne.901870103 (1979).
49 Balcita-Pedicino, J. J., Omelchenko, N., Bell, R. & Sesack, S. R. The inhibitory influence of the lateral habenula on midbrain dopamine cells: ultrastructural evidence for indirect mediation via the rostromedial mesopontine tegmental nucleus. *J Comp Neurol* **519**, 1143-1164, doi:10.1002/cne.22561 (2011).
50 Jhou, T. C., Fields, H. L., Baxter, M. G., Saper, C. B. & Holland, P. C. The rostromedial tegmental nucleus (RMTg), a GABAergic afferent to midbrain dopamine neurons, encodes aversive stimuli and inhibits motor responses. *Neuron* **61**, 786-800, doi:10.1016/j.neuron.2009.02.001 (2009).
51 Matsumoto, M. & Hikosaka, O. Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* **447**, 1111-1115, doi:10.1038/nature05860 (2007).
52 Matsumoto, M. & Hikosaka, O. Negative motivational control of saccadic eye movement by the lateral habenula. *Prog Brain Res* **171**, 399-402, doi:10.1016/S0079-6123(08)00658-4 (2008).
53 Johansen, J. P. & Fields, H. L. Glutamatergic activation of anterior cingulate cortex produces an aversive teaching signal. *Nat Neurosci* **7**, 398-403, doi:10.1038/nn1207 (2004).
54 Stamatakis, A. M. & Stuber, G. D. Activation of lateral habenula inputs to the ventral midbrain promotes behavioral avoidance. *Nat Neurosci* **15**, 1105-1107, doi:10.1038/nn.3145 (2012).
55 Matsumoto, M. & Hikosaka, O. Electrical stimulation of the primate lateral habenula suppresses saccadic eye movement through a learning mechanism. *PLoS One* **6**, e26701, doi:10.1371/journal.pone.0026701 (2011).
56 Christoph, G. R., Leonzio, R. J. & Wilcox, K. S. Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *J Neurosci* **6**, 613-619, doi:10.1523/JNEUROSCI.06-03-00613.1986 (1986).
57 Barrot, M. *et al.* Braking dopamine systems: a new GABA master structure for mesolimbic and nigrostriatal functions. *J Neurosci* **32**, 14094-14101, doi:10.1523/JNEUROSCI.3370-12.2012 (2012).
58 Vento, P. J. & Jhou, T. C. Bidirectional Valence Encoding in the Ventral Pallidum. *Neuron* **105**, 766-768, doi:10.1016/j.neuron.2020.02.017 (2020).
59 Ji, H. & Shepard, P. D. Lateral habenula stimulation inhibits rat midbrain dopamine neurons through a GABA(A) receptor-mediated mechanism. *J Neurosci* **27**, 6923-6930, doi:10.1523/JNEUROSCI.0958-07.2007 (2007).
60 Tian, J. & Uchida, N. Habenula Lesions Reveal that Multiple Mechanisms Underlie Dopamine Prediction Errors. *Neuron* **87**, 1304-1316, doi:10.1016/j.neuron.2015.08.028

(2015).
61 Malkova, L., Gaffan, D. & Murray, E. A. Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *J Neurosci* **17**, 6011-6020 (1997).
62 Mishkin, M., Vest, B., Waxler, M. & Rosvold, H. E. A re-examination of the effects of frontal lesions on object alternation. *Neuropsychologia* **7**, 357-363 (1969).
63 Thorpe, S., Rolls, E. & Maddison, S. The orbitofrontal cortex: neuronal activity in the behaving monkey. *Experimental brain research* **49**, 93-115 (1983).
64 Critchley, H. D. & Rolls, E. T. Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *Journal of neurophysiology* **75**, 1673-1686 (1996).
65 Tremblay, L. & Schultz, W. Relative reward preference in primate orbitofrontal cortex. *Nature* **398**, 704-708 (1999).
66 Padoa-Schioppa, C. Neurobiology of economic choice: a good-based model. *Annual review of neuroscience* **34**, 333 (2011).
67 Wallis, J. D. & Miller, E. K. Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience* **18**, 2069-2081 (2003).
68 Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223-226 (2006).
69 Sul, J. H., Kim, H., Huh, N., Lee, D. & Jung, M. W. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron* **66**, 449-460 (2010).
70 Rudebeck, P. H. & Murray, E. A. The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. *Neuron* **84**, 1143-1156 (2014).
71 Tang, H., Costa, V. D., Bartolo, R. & Averbeck, B. B. Differential coding of goals and actions in ventral and dorsal corticostriatal circuits during goal-directed behavior. *Cell Reports* **38**, 110198 (2022).
72 Rudebeck, P. H., Mitz, A. R., Chacko, R. V. & Murray, E. A. Effects of amygdala lesions on reward-value coding in orbital and medial prefrontal cortex. *Neuron* **80**, 1519-1531 (2013).
73 Grattan, L. & Glimcher, P. No evidence for spatial tuning in orbitofrontal cortex. *Soc. Neurosci. Meet. Plann* (2010).
74 Roesch, M. R. & Olson, C. R. Neuronal activity in primate orbitofrontal cortex reflects the value of time. *Journal of neurophysiology* **94**, 2457-2471 (2005).
75 Kennerley, S. W., Dahmubed, A. F., Lara, A. H. & Wallis, J. D. Neurons in the frontal lobe encode the value of multiple decision variables. *Journal of cognitive neuroscience* **21**, 1162-1178 (2009).
76 Morrison, S. E. & Salzman, C. D. The convergence of information about rewarding and aversive stimuli in single neurons. *Journal of Neuroscience* **29**, 11471-11483 (2009).
77 Amarante, L. M. & Laubach, M. Coherent theta activity in the medial and orbital frontal cortices encodes reward value. *Elife* **10**, e63372 (2021).
78 Rich, E. L. & Wallis, J. D. Decoding subjective decisions from orbitofrontal cortex. *Nature neuroscience* **19**, 973-980 (2016).
79 Price, J. L. Free will versus survival: brain systems that underlie intrinsic constraints on behavior. *Journal of Comparative Neurology* **493**, 132-139 (2005).
80 Sutton, R. S. & Barto, A. G. Introduction to reinforcement learning. (1998).

81 Kable, J. W. & Glimcher, P. W. The neural correlates of subjective value during intertemporal choice. *Nature neuroscience* **10**, 1625-1633 (2007).
82 Levy, I., Snell, J., Nelson, A. J., Rustichini, A. & Glimcher, P. W. Neural representation of subjective value under risk and ambiguity. *Journal of neurophysiology* **103**, 1036-1047 (2010).
83 Tom, S. M., Fox, C. R., Trepel, C. & Poldrack, R. A. The neural basis of loss aversion in decision-making under risk. *Science* **315**, 515-518 (2007).
84 Tang, H., Costa, V. D., Bartolo, R. & Averbeck, B. B. Differential coding of goals and actions in ventral and dorsal corticostriatal circuits during goal-directed behavior. *Cell Rep* **38**, 110198, doi:10.1016/j.celrep.2021.110198 (2022).
85 Cai, X. & Padoa-Schioppa, C. Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. *Neuron* **81**, 1140-1151, doi:10.1016/j.neuron.2014.01.008 (2014).
86 Tang, H., Bartolo, R. & Averbeck, B. B. Reward-related choices determine information timing and flow across macaque lateral prefrontal cortex. *Nat Commun* **12**, 894, doi:10.1038/s41467-021-20943-9 (2021).
87 Leon, M. I. & Shadlen, M. N. Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron* **24**, 415-425 (1999).
88 Barraclough, D. J., Conroy, M. L. & Lee, D. Prefrontal cortex and decision making in a mixed-strategy game. *Nat Neurosci* **7**, 404-410, doi:10.1038/nn1209 (2004).
89 Kim, S., Hwang, J. & Lee, D. Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron* **59**, 161-172, doi:10.1016/j.neuron.2008.05.010 (2008).
90 Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Dimensionality, information and learning in prefrontal cortex. *PLoS computational biology* **16**, e1007514 (2020).
91 Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585-590 (2013).
92 Seo, H. & Lee, D. Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J Neurosci* **27**, 8366-8377, doi:10.1523/JNEUROSCI.2369-07.2007 (2007).
93 McCoy, A. N. & Platt, M. L. Risk-sensitive neurons in macaque posterior cingulate cortex. *Nat Neurosci* **8**, 1220-1227, doi:10.1038/nn1523 (2005).
94 Fecteau, J. H. & Munoz, D. P. Exploring the consequences of the previous trial. *Nat Rev Neurosci* **4**, 435-443, doi:10.1038/nrn1114 (2003).
95 Seo, H., Barraclough, D. J. & Lee, D. Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *J Neurosci* **29**, 7278-7289, doi:10.1523/JNEUROSCI.1479-09.2009 (2009).
96 Seo, H. & Lee, D. Cortical mechanisms for reinforcement learning in competitive games. *Philos Trans R Soc Lond B Biol Sci* **363**, 3845-3857, doi:10.1098/rstb.2008.0158 (2008).
97 Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233-238, doi:10.1038/22268 (1999).
98 Dorris, M. C. & Glimcher, P. W. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* **44**, 365-378, doi:10.1016/j.neuron.2004.09.009 (2004).
99 Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782-1787, doi:10.1126/science.1094765 (2004).

100 Leathers, M. L. & Olson, C. R. In monkeys making value-based decisions, LIP neurons encode cue salience and not action value. *Science* **338**, 132-135, doi:10.1126/science.1226405 (2012).
101 Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci* **9**, 357-381, doi:10.1146/annurev.ne.09.030186.002041 (1986).
102 Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends in neurosciences* **12**, 366-375 (1989).
103 Parent, A. & Hazrati, L.-N. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain research reviews* **20**, 91-127 (1995).
104 Freeze, B. S., Kravitz, A. V., Hammack, N., Berke, J. D. & Kreitzer, A. C. Control of basal ganglia output by direct and indirect pathway projection neurons. *J Neurosci* **33**, 18531-18539, doi:10.1523/JNEUROSCI.1278-13.2013 (2013).
105 Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience* **9**, 357-381 (1986).
106 Bostan, A. C., Dum, R. P. & Strick, P. L. The basal ganglia communicate with the cerebellum. *Proceedings of the national academy of sciences* **107**, 8452-8456 (2010).
107 Bostan, A. C. & Strick, P. L. The basal ganglia and the cerebellum: nodes in an integrated network. *Nature Reviews Neuroscience* **19**, 338-350 (2018).
108 Houk, J. C., Davis, J. L. & Beiser, D. G. *Models of information processing in the basal ganglia*. (MIT press, 1995).
109 Haber, S. N., Fudge, J. L. & McFarland, N. R. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J Neurosci* **20**, 2369-2382, doi:10.1523/JNEUROSCI.20-06-02369.2000 (2000).
110 Lynd-Balta, E. & Haber, S. The organization of midbrain projections to the ventral striatum in the primate. *Neuroscience* **59**, 609-623 (1994).
111 Lynd-Balta, E. & Haber, S. The organization of midbrain projections to the striatum in the primate: sensorimotor-related striatum versus ventral striatum. *Neuroscience* **59**, 625-640 (1994).
112 O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452-454, doi:10.1126/science.1094285 (2004).
113 Cai, X., Kim, S. & Lee, D. Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron* **69**, 170-182, doi:10.1016/j.neuron.2010.11.041 (2011).
114 Hassani, O. K., Cromwell, H. C. & Schultz, W. Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. *J Neurophysiol* **85**, 2477-2489, doi:10.1152/jn.2001.85.6.2477 (2001).
115 Apicella, P., Legallet, E. & Trouche, E. Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Exp Brain Res* **116**, 456-466, doi:10.1007/pl00005773 (1997).
116 Aosaki, T. *et al.* Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *J Neurosci* **14**, 3969-3984, doi:10.1523/JNEUROSCI.14-06-03969.1994 (1994).
117 Falcone, R. *et al.* Temporal Coding of Reward Value in Monkey Ventral Striatal Tonically Active Neurons. *J Neurosci* **39**, 7539-7550, doi:10.1523/JNEUROSCI.0869-

19.2019 (2019).
118 Hikosaka, O., Sakamoto, M. & Usui, S. Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *J Neurophysiol* **61**, 814-832, doi:10.1152/jn.1989.61.4.814 (1989).
119 Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337-1340, doi:10.1126/science.1115270 (2005).
120 Santacruz, S. R., Rich, E. L., Wallis, J. D. & Carmena, J. M. Caudate Microstimulation Increases Value of Specific Choices. *Curr Biol* **27**, 3375-3383 e3373, doi:10.1016/j.cub.2017.09.051 (2017).
121 Rothenhoefer, K. M. *et al.* Effects of Ventral Striatum Lesions on Stimulus-Based versus Action-Based Reinforcement Learning. *J Neurosci* **37**, 6902-6914, doi:10.1523/JNEUROSCI.0631-17.2017 (2017).
122 Diederen, K. M., Spencer, T., Vestergaard, M. D., Fletcher, P. C. & Schultz, W. Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency. *Neuron* **90**, 1127-1138, doi:10.1016/j.neuron.2016.04.019 (2016).
123 Day, J. J., Roitman, M. F., Wightman, R. M. & Carelli, R. M. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat Neurosci* **10**, 1020-1028, doi:10.1038/nn1923 (2007).
124 Shidara, M., Aigner, T. G. & Richmond, B. J. Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *J Neurosci* **18**, 2613-2625, doi:10.1523/JNEUROSCI.18-07-02613.1998 (1998).
125 Simmons, J. M., Ravel, S., Shidara, M. & Richmond, B. J. A comparison of reward-contingent neuronal activity in monkey orbitofrontal cortex and ventral striatum: guiding actions toward rewards. *Ann N Y Acad Sci* **1121**, 376-394, doi:10.1196/annals.1401.028 (2007).
126 Strait, C. E., Sleezer, B. J. & Hayden, B. Y. Signatures of Value Comparison in Ventral Striatum Neurons. *PLoS Biol* **13**, e1002173, doi:10.1371/journal.pbio.1002173 (2015).
127 Costa, V. D., Mitz, A. R. & Averbeck, B. B. Subcortical Substrates of Explore-Exploit Decisions in Primates. *Neuron* **103**, 533-545 e535, doi:10.1016/j.neuron.2019.05.017 (2019).
128 Yamada, H., Imaizumi, Y. & Matsumoto, M. Neural Population Dynamics Underlying Expected Value Computation. *J Neurosci* **41**, 1684-1698, doi:10.1523/JNEUROSCI.1987-20.2020 (2021).
129 Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A. & Averbeck, B. B. Amygdala and Ventral Striatum Make Distinct Contributions to Reinforcement Learning. *Neuron* **92**, 505-517, doi:10.1016/j.neuron.2016.09.025 (2016).
130 Taswell, C. A., Costa, V. D., Murray, E. A. & Averbeck, B. B. Ventral striatum's role in learning from gains and losses. *Proc Natl Acad Sci U S A* **115**, E12398-E12406, doi:10.1073/pnas.1809833115 (2018).
131 Saga, Y., Hoshi, E. & Tremblay, L. Roles of Multiple Globus Pallidus Territories of Monkeys and Humans in Motivation, Cognition and Action: An Anatomical, Physiological and Pathophysiological Review. *Front Neuroanat* **11**, 30, doi:10.3389/fnana.2017.00030 (2017).
132 Middleton, F. A. & Strick, P. L. Basal ganglia output and cognition: evidence from

anatomical, behavioral, and clinical studies. *Brain Cogn* **42**, 183-200, doi:10.1006/brcg.1999.1099 (2000).
133 DeLong, M. R. Activity of pallidal neurons during movement. *J Neurophysiol* **34**, 414-427, doi:10.1152/jn.1971.34.3.414 (1971).
134 Desmurget, M. & Turner, R. S. Testing basal ganglia motor functions through reversible inactivations in the posterior internal globus pallidus. *J Neurophysiol* **99**, 1057-1076, doi:10.1152/jn.01010.2007 (2008).
135 Hong, S. & Hikosaka, O. The globus pallidus sends reward-related signals to the lateral habenula. *Neuron* **60**, 720-729, doi:10.1016/j.neuron.2008.09.035 (2008).
136 Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Neurophysiol* **104**, 1068-1076, doi:10.1152/jn.00158.2010 (2010).
137 Arkadir, D., Morris, G., Vaadia, E. & Bergman, H. Independent coding of movement direction and reward prediction by single pallidal neurons. *J Neurosci* **24**, 10047-10056, doi:10.1523/JNEUROSCI.2583-04.2004 (2004).
138 Smith, K. S., Tindell, A. J., Aldridge, J. W. & Berridge, K. C. Ventral pallidum roles in reward and motivation. *Behav Brain Res* **196**, 155-167, doi:10.1016/j.bbr.2008.09.038 (2009).
139 Kalivas, P. W. & Nakamura, M. Neural systems for behavioral activation and reward. *Curr Opin Neurobiol* **9**, 223-227, doi:10.1016/s0959-4388(99)80031-2 (1999).
140 Rajmohan, V. & Mohandas, E. The limbic system. *Indian J Psychiatry* **49**, 132-139, doi:10.4103/0019-5545.33264 (2007).
141 Brown, R. M., Crane, A. M. & Goldman, P. S. Regional distribution of monoamines in the cerebral cortex and subcortical structures of the rhesus monkey: concentrations and in vivo synthesis rates. *Brain Res* **168**, 133-150, doi:10.1016/0006-8993(79)90132-x (1979).
142 Cardinal, R. N., Parkinson, J. A., Hall, J. & Everitt, B. J. Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neurosci Biobehav Rev* **26**, 321-352, doi:10.1016/s0149-7634(02)00007-6 (2002).
143 LeDoux, J. E. Emotion circuits in the brain. *Annu Rev Neurosci* **23**, 155-184, doi:10.1146/annurev.neuro.23.1.155 (2000).
144 Baxter, M. G. & Murray, E. A. The amygdala and reward. *Nat Rev Neurosci* **3**, 563-573, doi:10.1038/nrn875 (2002).
145 Nishijo, H., Ono, T. & Nishino, H. Single neuron responses in amygdala of alert monkey during complex sensory stimulation with affective significance. *J Neurosci* **8**, 3570-3583, doi:10.1523/JNEUROSCI.08-10-03570.1988 (1988).
146 Belova, M. A., Paton, J. J. & Salzman, C. D. Moment-to-moment tracking of state value in the amygdala. *J Neurosci* **28**, 10023-10030, doi:10.1523/JNEUROSCI.1400-08.2008 (2008).
147 Morrison, S. E. & Salzman, C. D. Re-valuing the amygdala. *Curr Opin Neurobiol* **20**, 221-230, doi:10.1016/j.conb.2010.02.007 (2010).
148 Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439**, 865-870, doi:10.1038/nature04490 (2006).
149 Holland, P. C. & Gallagher, M. Amygdala-frontal interactions and reward expectancy. *Curr Opin Neurobiol* **14**, 148-155, doi:10.1016/j.conb.2004.03.007 (2004).
150 Rudebeck, P. H., Mitz, A. R., Chacko, R. V. & Murray, E. A. Effects of amygdala lesions

on reward-value coding in orbital and medial prefrontal cortex. *Neuron* **80**, 1519-1531, doi:10.1016/j.neuron.2013.09.036 (2013).
151 Ito, H., Takahashi, H., Arakawa, R., Takano, H. & Suhara, T. Normal database of dopaminergic neurotransmission system in human brain measured by positron emission tomography. *Neuroimage* **39**, 555-565, doi:10.1016/j.neuroimage.2007.09.011 (2008).
152 Merlo, E. *et al.* Amygdala Dopamine Receptors Are Required for the Destabilization of a Reconsolidating Appetitive Memory. *eNeuro* **2**, doi:10.1523/ENEURO.0024-14.2015 (2015).
153 Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Dopamine modulates novelty seeking behavior during decision making. *Behav Neurosci* **128**, 556-566, doi:10.1037/a0037128 (2014).
154 Marr, D. A theory of cerebellar cortex. *J Physiol* **202**, 437-470, doi:10.1113/jphysiol.1969.sp008820 (1969).
155 Taylor, J. A. & Ivry, R. B. Cerebellar and prefrontal cortex contributions to adaptation, strategies, and reinforcement learning. *Prog Brain Res* **210**, 217-253, doi:10.1016/B978-0-444-63356-9.00009-1 (2014).
156 Wagner, M. J., Kim, T. H., Savall, J., Schnitzer, M. J. & Luo, L. Cerebellar granule cells encode the expectation of reward. *Nature* **544**, 96-100, doi:10.1038/nature21726 (2017).
157 Garrison, J., Erdeniz, B. & Done, J. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci Biobehav Rev* **37**, 1297-1310, doi:10.1016/j.neubiorev.2013.03.023 (2013).
158 Bostan, A. C., Dum, R. P. & Strick, P. L. The basal ganglia communicate with the cerebellum. *Proc Natl Acad Sci U S A* **107**, 8452-8456, doi:10.1073/pnas.1000496107 (2010).
159 Bostan, A. C. & Strick, P. L. The basal ganglia and the cerebellum: nodes in an integrated network. *Nat Rev Neurosci* **19**, 338-350, doi:10.1038/s41583-018-0002-7 (2018).
160 Christopoulos, G. I., Tobler, P. N., Bossaerts, P., Dolan, R. J. & Schultz, W. Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *J Neurosci* **29**, 12574-12583, doi:10.1523/JNEUROSCI.2614-09.2009 (2009).
161 Pushkarskaya, H., Smithson, M., Joseph, J. E., Corbly, C. & Levy, I. Neural Correlates of Decision-Making Under Ambiguity and Conflict. *Front Behav Neurosci* **9**, 325, doi:10.3389/fnbeh.2015.00325 (2015).
162 Kahnt, T., Heinzle, J., Park, S. Q. & Haynes, J. D. Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. *Neuroimage* **56**, 709-715, doi:10.1016/j.neuroimage.2010.05.058 (2011).
163 Knoch, D. *et al.* Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. *J Neurosci* **26**, 6469-6472, doi:10.1523/JNEUROSCI.0804-06.2006 (2006).
164 Fecteau, S. *et al.* Activation of prefrontal cortex by transcranial direct current stimulation reduces appetite for risk during ambiguous decision making. *J Neurosci* **27**, 6212-6218, doi:10.1523/JNEUROSCI.0314-07.2007 (2007).
165 Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T. & Platt, M. L. Neural signatures of economic preferences for risk and ambiguity. *Neuron* **49**, 765-775, doi:10.1016/j.neuron.2006.01.024 (2006).
166 O'Neill, M. & Schultz, W. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron* **68**, 789-800,

doi:10.1016/j.neuron.2010.09.031 (2010).
167 Ogawa, M. *et al.* Risk-responsive orbitofrontal neurons track acquired salience. *Neuron* **77**, 251-258, doi:10.1016/j.neuron.2012.11.006 (2013).
168 White, J. K. & Monosov, I. E. Neurons in the primate dorsal striatum signal the uncertainty of object-reward associations. *Nat Commun* **7**, 12735, doi:10.1038/ncomms12735 (2016).
169 Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J Physiol* **148**, 574-591, doi:10.1113/jphysiol.1959.sp006308 (1959).
170 Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* **160**, 106-154, doi:10.1113/jphysiol.1962.sp006837 (1962).
171 Wurtz, R. H. Recounting the impact of Hubel and Wiesel. *J Physiol* **587**, 2817-2823, doi:10.1113/jphysiol.2009.170209 (2009).
172 Shuler, M. G. & Bear, M. F. Reward timing in the primary visual cortex. *Science* **311**, 1606-1609, doi:10.1126/science.1123513 (2006).
173 Stanisor, L., van der Togt, C., Pennartz, C. M. & Roelfsema, P. R. A unified selection signal for attention and reward in primary visual cortex. *Proc Natl Acad Sci U S A* **110**, 9136-9141, doi:10.1073/pnas.1300117110 (2013).
174 Lauter, J. L., Herscovitch, P., Formby, C. & Raichle, M. E. Tonotopic organization in human auditory cortex revealed by positron emission tomography. *Hear Res* **20**, 199-205, doi:10.1016/0378-5955(85)90024-3 (1985).
175 Merzenich, M. M., Knight, P. L. & Roth, G. L. Cochleotopic organization of primary auditory cortex in the cat. *Brain Res* **63**, 343-346, doi:10.1016/0006-8993(73)90101-7 (1973).
176 Recanzone, G. H., Schreiner, C. E., Sutter, M. L., Beitel, R. E. & Merzenich, M. M. Functional organization of spectral receptive fields in the primary auditory cortex of the owl monkey. *J Comp Neurol* **415**, 460-481, doi:10.1002/(sici)1096-9861(19991227)415:4<460::aid-cne4>3.0.co;2-f (1999).
177 Edeline, J. M. & Weinberger, N. M. Receptive field plasticity in the auditory cortex during frequency discrimination training: selective retuning independent of task difficulty. *Behav Neurosci* **107**, 82-103, doi:10.1037//0735-7044.107.1.82 (1993).
178 Edeline, J. M. Learning-induced physiological plasticity in the thalamo-cortical sensory systems: a critical evaluation of receptive field plasticity, map changes and their potential mechanisms. *Prog Neurobiol* **57**, 165-224, doi:10.1016/s0301-0082(98)00042-2 (1999).
179 Recanzone, G. H., Schreiner, C. E. & Merzenich, M. M. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J Neurosci* **13**, 87-103 (1993).
180 Wikman, P., Rinne, T. & Petkov, C. I. Reward cues readily direct monkeys' auditory performance resulting in broad auditory cortex modulation and interaction with sites along cholinergic and dopaminergic pathways. *Sci Rep* **9**, 3055, doi:10.1038/s41598-019-38833-y (2019).
181 Brunk, M. G. K. *et al.* Optogenetic stimulation of the VTA modulates a frequency-specific gain of thalamocortical inputs in infragranular layers of the auditory cortex. *Sci Rep* **9**, 20385, doi:10.1038/s41598-019-56926-6 (2019).
182 Penfield, W. & Boldrey, E. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* **60**, 389-443 (1937).

183 Fetz, E. E. & Cheney, P. D. Postspike facilitation of forelimb muscle activity by primate corticomotoneuronal cells. *J Neurophysiol* **44**, 751-772, doi:10.1152/jn.1980.44.4.751 (1980).
184 Georgopoulos, A. P., Kettner, R. E. & Schwartz, A. B. Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Journal of Neuroscience* **8**, 2928-2937 (1988).
185 Kalaska, J. F., Scott, S. H., Cisek, P. & Sergio, L. E. Cortical control of reaching movements. *Current opinion in neurobiology* **7**, 849-859 (1997).
186 Moran, D. W. & Schwartz, A. B. Motor cortical representation of speed and direction during reaching. *Journal of neurophysiology* **82**, 2676-2692 (1999).
187 Shmuelof, L. & Krakauer, J. W. Are we ready for a natural history of motor learning? *Neuron* **72**, 469-476, doi:10.1016/j.neuron.2011.10.017 (2011).
188 Musallam, S., Corneil, B. D., Greger, B., Scherberger, H. & Andersen, R. A. Cognitive control signals for neural prosthetics. *Science* **305**, 258-262, doi:DOI 10.1126/science.1097938 (2004).
189 Roesch, M. R. & Olson, C. R. Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *J Neurophysiol* **90**, 1766-1789, doi:10.1152/jn.00019.2003 (2003).
190 Tanaka, S. C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience* **7**, 887-893, doi:10.1038/nn1279 (2004).
191 Schultz, W. Behavioral theories and the neurophysiology of reward. *Annu Rev Psychol* **57**, 87-115, doi:10.1146/annurev.psych.56.091103.070229 (2006).
192 O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329-337, doi:10.1016/s0896-6273(03)00169-7 (2003).
193 Dayan, P. & Balleine, B. W. Reward, motivation, and reinforcement learning. *Neuron* **36**, 285-298, doi:10.1016/s0896-6273(02)00963-7 (2002).
194 Marsh, B. T., Tarigoppula, V. S. A., Chen, C. & Francis, J. T. Toward an autonomous brain machine interface: integrating sensorimotor reward modulation and reinforcement learning. *Journal of Neuroscience* **35**, 7374-7387 (2015).
195 Ramkumar, P., Dekleva, B., Cooler, S., Miller, L. & Kording, K. Premotor and Motor Cortices Encode Reward. *PLoS One* **11**, e0160851, doi:10.1371/journal.pone.0160851 (2016).
196 Thabit, M. N. *et al.* Momentary reward induce changes in excitability of primary motor cortex. *Clinical Neurophysiology* **122**, 1764-1770 (2011).
197 An, J., Yadav, T., Hessburg, J. P. & Francis, J. T. Reward expectation modulates local field potentials, spiking activity and spike-field coherence in the primary motor cortex. *eneuro* **6** (2019).
198 Richfield, E. K., Young, A. B. & Penney, J. B. Comparative Distributions of Dopamine D-1 and D-2 Receptors in the Cerebral-Cortex of Rats, Cats, and Monkeys. *Journal of Comparative Neurology* **286**, 409-426, doi:DOI 10.1002/cne.902860402 (1989).
199 Lewis, D., Campbell, M., Foote, S., Goldstein, M. & Morrison, J. The distribution of tyrosine hydroxylase-immunoreactive fibers in primate neocortex is widespread but regionally specific. *Journal of Neuroscience* **7**, 279-290 (1987).
200 Descarries, L., Lemay, B., Doucet, G. & Berger, B. Regional and laminar density of the

dopamine innervation in adult rat cerebral cortex. *Neuroscience* **21**, 807-824 (1987).
201 Luft, A. R. & Schwarz, S. Dopaminergic signals in primary motor cortex. *International Journal of Developmental Neuroscience* **27**, 415-421 (2009).
202 Doyon, J. Motor sequence learning and movement disorders. *Current opinion in neurology* **21**, 478-483 (2008).
203 Molina-Luna, K. *et al.* Dopamine in motor cortex is necessary for skill learning and synaptic plasticity. *PloS one* **4**, e7082 (2009).
204 Hosp, J. A., Pekanovic, A., Rioult-Pedotti, M. S. & Luft, A. R. Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning. *Journal of Neuroscience* **31**, 2481-2487 (2011).
205 O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research* (1971).
206 Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801-806 (2005).
207 Kropff, E., Carmichael, J. E., Moser, M.-B. & Moser, E. I. Speed cells in the medial entorhinal cortex. *Nature* **523**, 419-424 (2015).
208 Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B. & Moser, E. I. Representation of geometric borders in the entorhinal cortex. *Science* **322**, 1865-1868 (2008).
209 Sargolini, F. *et al.* Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **312**, 758-762 (2006).
210 Butler, W. N., Hardcastle, K. & Giocomo, L. M. Remembered reward locations restructure entorhinal spatial maps. *Science* **363**, 1447-1452, doi:10.1126/science.aav5297 (2019).
211 Lee, J. Y. *et al.* Dopamine facilitates associative memory encoding in the entorhinal cortex. *Nature* **598**, 321-326, doi:10.1038/s41586-021-03948-8 (2021).
212 Watanabe, T. & Niki, H. Hippocampal unit activity and delayed response in the monkey. *Brain research* **325**, 241-254 (1985).
213 Gauthier, J. L. & Tank, D. W. A Dedicated Population for Reward Coding in the Hippocampus. *Neuron* **99**, 179-193 e177, doi:10.1016/j.neuron.2018.06.008 (2018).
214 Ljungberg, T., Apicella, P. & Schultz, W. Responses of monkey dopamine neurons during learning of behavioral reactions. *J Neurophysiol* **67**, 145-163, doi:10.1152/jn.1992.67.1.145 (1992).
215 Mirenowicz, J. & Schultz, W. Importance of unpredictability for reward responses in primate dopamine neurons. *J Neurophysiol* **72**, 1024-1027 (1994).
216 Mirenowicz, J. & Schultz, W. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* **379**, 449-451, doi:10.1038/379449a0 (1996).
217 Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43-48, doi:10.1038/35083500 (2001).
218 Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science* **299**, 1898-1902 (2003).
219 Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y. & Hikosaka, O. Dopamine neurons can represent context-dependent prediction error. *Neuron* **41**, 269-280 (2004).
220 Stauffer, W. R. The biological and behavioral computations that influence dopamine responses. *Current Opinion in Neurobiology* **49**, 123-131, doi:https://doi.org/10.1016/j.conb.2018.02.005 (2018).

221 Enomoto, K. *et al.* Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proc Natl Acad Sci U S A* **108**, 15462-15467, doi:10.1073/pnas.1014457108 (2011).
222 Sutton, R. & Barto, A. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).
223 d'Acremont, M. & Bossaerts, P. Neural Mechanisms Behind Identification of Leptokurtic Noise and Adaptive Behavioral Response. *Cerebral Cortex* **26**, 1818-1830, doi:10.1093/cercor/bhw013 (2016).
224 Diederen, K. M. J. & Schultz, W. Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology* **114**, 1628-1640, doi:10.1152/jn.00483.2015 (2015).
225 Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J Neurosci* **30**, 12366-12378, doi:10.1523/JNEUROSCI.0822-10.2010 (2010).
226 Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat Neurosci* **13**, 1292-1298, doi:10.1038/nn.2635 (2010).
227 Milosavljevic, M., Malmaud, J., Huth, A., Koch, C. & Rangel, A. The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgm Decis Mak* **5**, 437-449 (2010).
228 Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529-533, doi:10.1038/nature14236 (2015).
229 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489, doi:10.1038/nature16961 (2016).
230 Bellemare, M. G., Dabney, W. & Munos, R. in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* 449-458 (JMLR.org, Sydney, NSW, Australia, 2017).
231 Dabney, W. *et al.* A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671-675, doi:10.1038/s41586-019-1924-6 (2020).
232 Gershman, S. J. A Unifying Probabilistic View of Associative Learning. *PLoS Comput Biol* **11**, e1004567, doi:10.1371/journal.pcbi.1004567 (2015).
233 Loe, P. R., Whitsel, B. L., Dreyer, D. A. & Metz, C. B. Body representation in ventrobasal thalamus of macaque: a single-unit analysis. *J Neurophysiol* **40**, 1339-1355, doi:10.1152/jn.1977.40.6.1339 (1977).
234 Guyenet, P. G. & Aghajanian, G. K. Antidromic identification of dopaminergic and other output neurons of the rat substantia nigra. *Brain Res* **150**, 69-84, doi:10.1016/0006-8993(78)90654-6 (1978).
235 Batista, A. P. *et al.* Cortical neural prosthesis performance improves when eye position is monitored. *IEEE Trans Neural Syst Rehabil Eng* **16**, 24-31, doi:10.1109/TNSRE.2007.906958 (2008).
236 Van Slooten, J. C., Jahfari, S., Knapen, T. & Theeuwes, J. How pupil responses track value-based decision-making during and after reinforcement learning. *PLOS Computational Biology* **14**, e1006632, doi:10.1371/journal.pcbi.1006632 (2018).
237 de Hollander, G. & Knapen, T. *nideconv*, <https://nideconv.readthedocs.io/en/latest/> (2017).
238 Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*. (Random House Publishing Group, 2007).

239 Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* **16**, 1936-1947 (1996).
240 Lak, A., Stauffer, W. R. & Schultz, W. Dopamine neurons learn relative chosen value from probabilistic rewards. *Elife* **5**, doi:10.7554/eLife.18044 (2016).
241 Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837-841, doi:10.1038/nature08028 (2009).
242 Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci* **9**, 1057-1063, doi:10.1038/nn1743 (2006).
243 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489, doi:10.1038/nature16961 (2016).
244 Babayan, B. M., Uchida, N. & Gershman, S. J. Belief state representation in the dopamine system. *Nat Commun* **9**, 1891, doi:10.1038/s41467-018-04397-0 (2018).
245 Morrens, J., Aydin, Ç., Janse van Rensburg, A., Esquivelzeta Rabell, J. & Haesler, S. Cue-Evoked Dopamine Promotes Conditioned Responding during Learning. *Neuron* **106**, 142-153.e147, doi:10.1016/j.neuron.2020.01.012 (2020).
246 Preuschoff, K., Marius't Hart, B. & Einhauser, W. Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Frontiers in Neuroscience* **5**, doi:10.3389/fnins.2011.00115 (2011).
247 Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiol Rev* **95**, 853-951, doi:10.1152/physrev.00023.2014 (2015).
248 Vijayraghavan, S., Wang, M., Birnbaum, S. G., Williams, G. V. & Arnsten, A. F. Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. *Nat Neurosci* **10**, 376-384, doi:10.1038/nn1846 (2007).
249 Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect marginal utility. *Current Biololgy* **24**, 2491-2500, doi:10.1016/j.cub.2014.08.064 (2014).
250 Fox, C. R. & Tversky, A. Ambiguity aversion and comparative ignorance. *Q. J. Econ* **110**, 585-603 (1995).
251 Health, C. & Tversky, A. Preference and Belief: Ambiguity and Competence in Choice Under Uncertainty. *Journal of Risk and Uncertainty* **IV**, 5-28 (1991).
252 Camerer, C. & Weber, M. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of risk and uncertainty* **5**, 325-370 (1992).
253 Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C. F. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**, 1680-1683 (2005).
254 Hayden, B. Y., Heilbronner, S. R. & Platt, M. L. Ambiguity aversion in rhesus macaques. *Front Neurosci* **4**, doi:10.3389/fnins.2010.00166 (2010).
255 Jia, R., Furlong, E., Gao, S., Santos, L. R. & Levy, I. Learning about the Ellsberg Paradox reduces, but does not abolish, ambiguity aversion. *PLoS One* **15**, e0228782, doi:10.1371/journal.pone.0228782 (2020).
256 Stauffer, W. R., Lak, A., Bossaerts, P. & Schultz, W. Economic choices reveal probability distortion in macaque monkeys. *Journal of Neuroscience* **35**, 3146-3154, doi:10.1523/JNEUROSCI.3653-14.2015 (2015).
257 Vincent, P., Parr, T., Benrimoh, D. & Friston, K. J. With an eye on uncertainty: Modelling pupillary responses to environmental volatility. *PLoS Comput Biol* **15**,

e1007126, doi:10.1371/journal.pcbi.1007126 (2019).
258 Filipowicz, A. L., Glaze, C. M., Kable, J. W. & Gold, J. I. Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. *Elife* **9**, doi:10.7554/eLife.57872 (2020).
259 Colizoli, O., de Gee, J. W., Urai, A. E. & Donner, T. H. Task-evoked pupil responses reflect internal belief states. *Sci Rep* **8**, 13702, doi:10.1038/s41598-018-31985-3 (2018).
260 Mason, J. W. A review of psychoendocrine research on the sympathetic-adrenal medullary system. *Psychosom Med* **30**, Suppl:631-653, doi:10.1097/00006842-196809000-00022 (1968).
261 Mason, J. W. A review of psychoendocrine research on the pituitary-adrenal cortical system. *Psychosom Med* **30**, Suppl:576-607 (1968).
262 Monat, A., Averill, J. R. & Lazarus, R. S. Anticipatory stress and coping reactions under various conditions of uncertainty. *J Pers Soc Psychol* **24**, 237-253, doi:10.1037/h0033297 (1972).
263 Poe, G. R. *et al.* Locus coeruleus: a new look at the blue spot. *Nat Rev Neurosci* **21**, 644-659, doi:10.1038/s41583-020-0360-9 (2020).
264 Arnsten, A. F. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat Rev Neurosci* **10**, 410-422, doi:10.1038/nrn2648 (2009).
265 Jansen, A. S., Nguyen, X. V., Karpitskiy, V., Mettenleiter, T. C. & Loewy, A. D. Central command neurons of the sympathetic nervous system: basis of the fight-or-flight response. *Science* **270**, 644-646, doi:10.1126/science.270.5236.644 (1995).
266 Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience* **15**, 1040-1046, doi:10.1038/nn.3130 (2012).
267 Rothenhoefer, K. M., Hong, T., Alikaya, A. & Stauffer, W. R. Rare rewards amplify dopamine responses. *Nat Neurosci* **24**, 465-469, doi:10.1038/s41593-021-00807-7 (2021).
268 Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).
269 Stauffer, W. R., Lak, A., Bossaerts, P. & Schultz, W. Economic choices reveal probability distortion in macaque monkeys. *J Neurosci* **35**, 3146-3154, doi:10.1523/JNEUROSCI.3653-14.2015 (2015).
270 Preuschoff, K., t Hart, B. M. & Einhauser, W. Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Front Neurosci* **5**, 115, doi:10.3389/fnins.2011.00115 (2011).
271 Payzan-LeNestour, E. & Bossaerts, P. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput Biol* **7**, e1001048, doi:10.1371/journal.pcbi.1001048 (2011).
272 Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681-692, doi:10.1016/j.neuron.2005.04.026 (2005).
273 Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat Neurosci* **15**, 1040-1046, doi:10.1038/nn.3130 (2012).
274 Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C. F. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**, 1680-1683, doi:10.1126/science.1115327 (2005).
275 DiFiglia, M., Pasik, P. & Pasik, T. A Golgi study of neuronal types in the neostriatum of monkeys. *Brain Res* **114**, 245-256, doi:10.1016/0006-8993(76)90669-7 (1976).

276 Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends Neurosci* **12**, 366-375 (1989).
277 DeLong, M. R. & Wichmann, T. Basal Ganglia Circuits as Targets for Neuromodulation in Parkinson Disease. *JAMA Neurol* **72**, 1354-1360, doi:10.1001/jamaneurol.2015.2397 (2015).
278 Graybiel, A. M. & Ragsdale, C. W., Jr. Histochemically distinct compartments in the striatum of human, monkeys, and cat demonstrated by acetylthiocholinesterase staining. *Proc Natl Acad Sci U S A* **75**, 5723-5726 (1978).
279 Hong, S. *et al.* Predominant Striatal Input to the Lateral Habenula in Macaques Comes from Striosomes. *Curr Biol* **29**, 51-61 e55, doi:10.1016/j.cub.2018.11.008 (2019).
280 Gerfen, C. R. The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature* **311**, 461-464, doi:10.1038/311461a0 (1984).
281 Haber, S. N. & McFarland, N. R. The concept of the ventral striatum in nonhuman primates. *Ann N Y Acad Sci* **877**, 33-48, doi:10.1111/j.1749-6632.1999.tb09259.x (1999).
282 Heimer, L. & Wilson, R. in *Golgi centennial symposium proceedings.* 173-193 (Raven).
283 Voorn, P., Brady, L. S., Berendse, H. W. & Richfield, E. K. Densitometrical analysis of opioid receptor ligand binding in the human striatum--I. Distribution of mu opioid receptor defines shell and core of the ventral striatum. *Neuroscience* **75**, 777-792, doi:10.1016/0306-4522(96)00271-0 (1996).
284 Daunais, J. B. *et al.* Functional and anatomical localization of mu opioid receptors in the striatum, amygdala, and extended amygdala of the nonhuman primate. *J Comp Neurol* **433**, 471-485 (2001).
285 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382, doi:10.1038/nmeth.1315 (2009).
286 Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *Cell Rep* **16**, 1126-1137, doi:10.1016/j.celrep.2016.06.059 (2016).
287 Munoz-Manchado, A. B. *et al.* Diversity of Interneurons in the Dorsal Striatum Revealed by Single-Cell RNA Sequencing and PatchSeq. *Cell Rep* **24**, 2179-2190 e2177, doi:10.1016/j.celrep.2018.07.053 (2018).
288 Stanley, G., Gokce, O., Malenka, R. C., Sudhof, T. C. & Quake, S. R. Continuous and Discrete Neuron Types of the Adult Murine Striatum. *Neuron* **105**, 688-699 e688, doi:10.1016/j.neuron.2019.11.004 (2020).
289 Martin, A. *et al.* A Spatiomolecular Map of the Striatum. *Cell Rep* **29**, 4320-4333 e4325, doi:10.1016/j.celrep.2019.11.096 (2019).
290 Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-1030 e1016, doi:10.1016/j.cell.2018.07.028 (2018).
291 Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* **586**, 262-269, doi:10.1038/s41586-020-2781-z (2020).
292 Savell, K. E. *et al.* A dopamine-induced gene expression signature regulates neuronal function and cocaine response. *Sci Adv* **6**, eaba4221, doi:10.1126/sciadv.aba4221 (2020).
293 Lee, H. *et al.* Cell Type-Specific Transcriptomics Reveals that Mutant Huntingtin Leads to Mitochondrial RNA Release and Neuronal Innate Immune Activation. *Neuron* **107**, 891-908 e898, doi:10.1016/j.neuron.2020.06.021 (2020).
294 Izpisua Belmonte, J. C. *et al.* Brains, genes, and primates. *Neuron* **86**, 617-631, doi:10.1016/j.neuron.2015.03.021 (2015).
295 Davenport, A. T., Grant, K. A., Szeliga, K. T., Friedman, D. P. & Daunais, J. B.

Standardized method for the harvest of nonhuman primate tissue optimized for multiple modes of analyses. *Cell Tissue Bank* **15**, 99-110, doi:10.1007/s10561-013-9380-2 (2014).
296 Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* **14**, 955-958, doi:10.1038/nmeth.4407 (2017).
297 Warren, W. C. *et al.* Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**, doi:10.1126/science.abc6617 (2020).
298 Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, doi:10.1093/gigascience/giaa151 (2020).
299 McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e324, doi:10.1016/j.cels.2019.03.003 (2019).
300 Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61-68, doi:10.1038/s41586-019-1506-7 (2019).
301 Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78, doi:10.1038/s41586-018-0654-5 (2018).
302 Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014 e1022, doi:10.1016/j.cell.2018.06.021 (2018).
303 Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-11947, doi:10.1523/JNEUROSCI.1860-14.2014 (2014).
304 Grillner, S. & Robertson, B. The Basal Ganglia Over 500 Million Years. *Curr Biol* **26**, R1088-R1100, doi:10.1016/j.cub.2016.06.041 (2016).
305 Khrameeva, E. *et al.* Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res* **30**, 776-789, doi:10.1101/gr.256958.119 (2020).
306 Yin, S. *et al.* Transcriptomic and open chromatin atlas of high-resolution anatomical regions in the rhesus macaque brain. *Nat Commun* **11**, 474, doi:10.1038/s41467-020-14368-z (2020).
307 Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198, doi:10.1093/nar/gkz369 (2019).
308 Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**, 59, doi:10.1186/s13059-019-1663-x (2019).
309 Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* **20**, 295, doi:10.1186/s13059-019-1861-6 (2019).
310 Bor, J., Moscoe, E., Mutevedzi, P., Newell, M. L. & Barnighausen, T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology* **25**, 729-737, doi:10.1097/EDE.0000000000000138 (2014).
311 Zhang, J. M., Kamath, G. M. & Tse, D. N. Valid Post-clustering Differential Analysis for Single-Cell RNA-Seq. *Cell Syst* **9**, 383-392 e386, doi:10.1016/j.cels.2019.07.012 (2019).
312 Miao, Z. *et al.* Putative cell type discovery from single-cell gene expression data. *Nat Methods* **17**, 621-628, doi:10.1038/s41592-020-0825-9 (2020).
313 Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107

e1017, doi:10.1016/j.cell.2018.02.001 (2018).
314 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671-675, doi:10.1038/nmeth.2089 (2012).
315 Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7**, R100, doi:10.1186/gb-2006-7-10-r100 (2006).
316 Arlotta, P., Molyneaux, B. J., Jabaudon, D., Yoshida, Y. & Macklis, J. D. Ctip2 controls the differentiation of medium spiny neurons and the establishment of the cellular architecture of the striatum. *J Neurosci* **28**, 622-632, doi:10.1523/JNEUROSCI.2986-07.2008 (2008).
317 Xie, Z. *et al.* Cellular and subcellular localization of PDE10A, a striatum-enriched phosphodiesterase. *Neuroscience* **139**, 597-607, doi:10.1016/j.neuroscience.2005.12.042 (2006).
318 Crittenden, J. R. & Graybiel, A. M. Basal Ganglia disorders associated with imbalances in the striatal striosome and matrix compartments. *Front Neuroanat* **5**, 59, doi:10.3389/fnana.2011.00059 (2011).
319 Smith, J. B. *et al.* Genetic-Based Dissection Unveils the Inputs and Outputs of Striatal Patch and Matrix Compartments. *Neuron* **91**, 1069-1084, doi:10.1016/j.neuron.2016.07.046 (2016).
320 Mohammadi, S., Davila-Velderrain, J. & Kellis, M. A multiresolution framework to characterize single-cell state landscapes. *Nature Communications* **11**, 5399, doi:10.1038/s41467-020-18416-6 (2020).
321 Mohammadi, S., Ravindra, V., Gleich, D. F. & Grama, A. A geometric approach to characterize the functional identity of single cells. *Nature Communications* **9**, 1516, doi:10.1038/s41467-018-03933-2 (2018).
322 Wei, S. C. *et al.* Negative Co-stimulation Constrains T Cell Differentiation by Imposing Boundaries on Possible Cell States. *Immunity* **50**, 1084-1098.e1010, doi:https://doi.org/10.1016/j.immuni.2019.03.004 (2019).
323 Puighermanal, E. *et al.* Functional and molecular heterogeneity of D2R neurons along dorsal ventral axis in the striatum. *Nature Communications* **11**, 1957, doi:10.1038/s41467-020-15716-9 (2020).
324 Mikula, S., Parrish, S. K., Trimmer, J. S. & Jones, E. G. Complete 3D visualization of primate striosomes by KChIP1 immunostaining. *J Comp Neurol* **514**, 507-517, doi:10.1002/cne.22051 (2009).
325 Meredith, G. E., Pattiselanno, A., Groenewegen, H. J. & Haber, S. N. Shell and core in monkey and human nucleus accumbens identified with antibodies to calbindin-D28k. *J Comp Neurol* **365**, 628-639, doi:10.1002/(sici)1096-9861(19960219)365:4<628::Aid-cne9>3.0.Co;2-6 (1996).
326 Prensa, L., Richard, S. & Parent, A. Chemical anatomy of the human ventral striatum and adjacent basal forebrain structures. *J Comp Neurol* **460**, 345-367, doi:10.1002/cne.10627 (2003).
327 Heimer, L. Basal forebrain in the context of schizophrenia. *Brain Res Brain Res Rev* **31**, 205-235, doi:10.1016/s0165-0173(99)00039-9 (2000).
328 Calabrese, E. *et al.* A diffusion tensor MRI atlas of the postmortem rhesus macaque brain. *Neuroimage* **117**, 408-416, doi:10.1016/j.neuroimage.2015.05.072 (2015).
329 Bakker, R., Tiesinga, P. & Kotter, R. The Scalable Brain Atlas: Instant Web-Based

Access to Public Brain Atlases and Related Content. *Neuroinformatics* **13**, 353-366, doi:10.1007/s12021-014-9258-x (2015).
330 Meyer, G., Gonzalez-Hernandez, T., Carrillo-Padilla, F. & Ferres-Torres, R. Aggregations of granule cells in the basal forebrain (islands of Calleja): Golgi and cytoarchitectonic study in different mammals, including man. *J Comp Neurol* **284**, 405-428, doi:10.1002/cne.902840308 (1989).
331 Furuta, T. & Kaneko, T. Third pathway in the cortico-basal ganglia loop: Neurokinin B-producing striatal neurons modulate cortical activity via striato-innominato-cortical projection. *Neuroscience Research* **54**, 1-10, doi:https://doi.org/10.1016/j.neures.2005.10.002 (2006).
332 Huerta-Ocampo, I., Mena-Segovia, J. & Bolam, J. P. Convergence of cortical and thalamic input to direct and indirect pathway medium spiny neurons in the striatum. *Brain Struct Funct* **219**, 1787-1800, doi:10.1007/s00429-013-0601-z (2014).
333 Eblen, F. & Graybiel, A. M. Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. *J Neurosci* **15**, 5999-6013, doi:10.1523/JNEUROSCI.15-09-05999.1995 (1995).
334 Fujiyama, F. *et al.* Exclusive and common targets of neostriatofugal projections of rat striosome neurons: a single neuron-tracing study using a viral vector. *European Journal of Neuroscience* **33**, 668-677, doi:https://doi.org/10.1111/j.1460-9568.2010.07564.x (2011).
335 Crittenden, J. R. *et al.* Striosome–dendron bouquets highlight a unique striatonigral circuit targeting dopamine-containing neurons. *Proceedings of the National Academy of Sciences* **113**, 11318, doi:10.1073/pnas.1613337113 (2016).
336 Friedman, A. *et al.* A Corticostriatal Path Targeting Striosomes Controls Decision-Making under Conflict. *Cell* **161**, 1320-1333, doi:https://doi.org/10.1016/j.cell.2015.04.049 (2015).
337 Friedman, A. *et al.* Striosomes Mediate Value-Based Learning Vulnerable in Age and a Huntington's Disease Model. *Cell* **183**, 918-934.e949, doi:https://doi.org/10.1016/j.cell.2020.09.060 (2020).
338 Mogenson, G. J., Jones, D. L. & Yim, C. Y. From motivation to action: functional interface between the limbic system and the motor system. *Prog Neurobiol* **14**, 69-97, doi:10.1016/0301-0082(80)90018-0 (1980).
339 Floresco, S. B. The nucleus accumbens: an interface between cognition, emotion, and action. *Annu Rev Psychol* **66**, 25-52, doi:10.1146/annurev-psych-010213-115159 (2015).
340 Furuta, T., Mori, T., Lee, T. & Kaneko, T. Third group of neostriatofugal neurons: Neurokinin B-producing neurons that send axons predominantly to the substantia innominata. *Journal of Comparative Neurology* **426**, 279-296, doi:https://doi.org/10.1002/1096-9861(20001016)426:2<279::AID-CNE9>3.0.CO;2-F (2000).
341 Zhou, L., Furuta, T. & Kaneko, T. Neurokinin B-producing projection neurons in the lateral stripe of the striatum and cell clusters of the accumbens nucleus in the rat. *Journal of Comparative Neurology* **480**, 143-161, doi:https://doi.org/10.1002/cne.20331 (2004).
342 Xiao, X. *et al.* A Genetically Defined Compartmentalized Striatal Direct Pathway for Negative Reinforcement. *Cell* **183**, 211-227 e220, doi:10.1016/j.cell.2020.08.032 (2020).
343 Berridge, K. C. & Kringelbach, M. L. Pleasure systems in the brain. *Neuron* **86**, 646-664, doi:10.1016/j.neuron.2015.02.018 (2015).

344 Grillner, S., Robertson, B. & Stephenson-Jones, M. The evolutionary origin of the vertebrate basal ganglia and its role in action selection. *The Journal of Physiology* **591**, 5425-5431, doi:https://doi.org/10.1113/jphysiol.2012.246660 (2013).
345 Lecumberri, A., Lopez-Janeiro, A., Corral-Domenge, C. & Bernacer, J. Neuronal density and proportion of interneurons in the associative, sensorimotor and limbic human striatum. *Brain Struct Funct* **223**, 1615-1625, doi:10.1007/s00429-017-1579-8 (2018).
346 Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* **21**, 12, doi:10.1186/s13059-019-1850-9 (2020).
347 Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology* **38**, 747-755, doi:10.1038/s41587-020-0469-4 (2020).
348 O'Leary, T. P. *et al.* Extensive and spatially variable within-cell-type heterogeneity across the basolateral amygdala. *eLife* **9**, e59003, doi:10.7554/eLife.59003 (2020).
349 Ramón y Cajal, S. *Histologie du systËme nerveux de l'homme & des vertÈbrÈs*. (Maloine, 1909).
350 Stein-O'Brien, G. L. *et al.* Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Systems* **8**, 395-411.e398, doi:https://doi.org/10.1016/j.cels.2019.04.004 (2019).
351 Di Bella, D. J. *et al.* Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature*, doi:10.1038/s41586-021-03670-5 (2021).
352 Bakken, T. E. *et al.* Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *bioRxiv*, 2020.2003.2031.016972, doi:10.1101/2020.03.31.016972 (2020).
353 Evrard, H. C., Forro, T. & Logothetis, N. K. Von Economo neurons in the anterior insula of the macaque monkey. *Neuron* **74**, 482-489, doi:10.1016/j.neuron.2012.03.003 (2012).
354 Genest, W., Stauffer, W. R. & Schultz, W. Utility functions predict variance and skewness risk preferences in monkeys. *Proc Natl Acad Sci U S A* **113**, 8402-8407, doi:10.1073/pnas.1602217113 (2016).
355 Haroush, K. & Williams, Z. M. Neuronal prediction of opponent's behavior during cooperative social interchange in primates. *Cell* **160**, 1233-1245, doi:10.1016/j.cell.2015.01.045 (2015).
356 Hart, A. S., Rutledge, R. B., Glimcher, P. W. & Phillips, P. E. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience* **34**, 698-704 (2014).
357 Sara, S. J. The locus coeruleus and noradrenergic modulation of cognition. *Nat Rev Neurosci* **10**, 211-223, doi:10.1038/nrn2573 (2009).
358 Weinshenker, D. & Schroeder, J. P. There and back again: a tale of norepinephrine and drug addiction. *Neuropsychopharmacology* **32**, 1433-1451, doi:10.1038/sj.npp.1301263 (2007).
359 Mathot, S. Pupillometry: Psychology, Physiology, and Function. *J Cogn* **1**, 16, doi:10.5334/joc.18 (2018).
360 de Gee, J. W. *et al.* Dynamic modulation of decision biases by brainstem arousal systems. *Elife* **6**, doi:10.7554/eLife.23232 (2017).
361 Asan, E. Ultrastructural features of tyrosine-hydroxylase-immunoreactive afferents and their targets in the rat amygdala. *Cell Tissue Res* **288**, 449-469, doi:10.1007/s004410050832 (1997).
362 Takahashi, H. *et al.* Contribution of dopamine D1 and D2 receptors to amygdala activity

in human. *J Neurosci* **30**, 3043-3047, doi:10.1523/JNEUROSCI.5689-09.2010 (2010).
363 Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Curr Biol* **27**, 821-832, doi:10.1016/j.cub.2017.02.026 (2017).
364 Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759-764, doi:10.1126/science.1169405 (2009).
365 Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227-231, doi:10.1038/nature07200 (2008).
366 Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* **88**, 78-92, doi:10.1016/j.neuron.2015.09.039 (2015).
367 Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat Commun* **8**, 14637, doi:10.1038/ncomms14637 (2017).
368 Pulford, B. D. Is luck on my side? optimism, pessimism, and ambiguity aversion. *Quarterly Journal of Experimental Psychology* **62**, 1079-1087 (2009).
369 Amemori, K. & Graybiel, A. M. Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. *Nat Neurosci* **15**, 776-785, doi:10.1038/nn.3088 (2012).
370 Aupperle, R. L. & Paulus, M. P. Neural systems underlying approach and avoidance in anxiety disorders. *Dialogues Clin Neurosci* **12**, 517-531 (2010).
371 Szanto, K. *et al.* Decision-making competence and attempted suicide. *J Clin Psychiatry* **76**, e1590-1597, doi:10.4088/JCP.15m09778 (2015).
372 Li, P., Snyder, G. L. & Vanover, K. E. Dopamine Targeting Drugs for the Treatment of Schizophrenia: Past, Present and Future. *Curr Top Med Chem* **16**, 3385-3403, doi:10.2174/1568026616666160608084834 (2016).
373 Bromberg-Martin, E. S. & Hikosaka, O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* **63**, 119-126 (2009).
374 Kobayashi, S. & Schultz, W. Influence of reward delays on responses of dopamine neurons. *Journal of neuroscience* **28**, 7837-7846 (2008).
375 Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect marginal utility. *Current biology* **24**, 2491-2500 (2014).
376 Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nature neuroscience* **11**, 966-973 (2008).
377 Rothenhoefer, K. M., Hong, T., Alikaya, A. & Stauffer, W. R. Rare rewards amplify dopamine responses. *Nature neuroscience* **24**, 465-469 (2021).
378 Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042-1045 (2006).
379 Parkinson, J. *et al.* Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behavioural brain research* **137**, 149-163 (2002).
380 Czernecki, V. *et al.* Motivation, reward, and Parkinson's disease: influence of dopatherapy. *Neuropsychologia* **40**, 2257-2267 (2002).
381 Fischbach, S. & Janak, P. H. Decreases in cued reward seeking after reward-paired

inhibition of mesolimbic dopamine. *Neuroscience* **412**, 259-269 (2019).
382 Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and learning. *Nature neuroscience* **16**, 966-973 (2013).
383 Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of neuroscience* **16**, 1936-1947 (1996).
384 Moore, B. D. t. & Freeman, R. D. Development of orientation tuning in simple cells of primary visual cortex. *J Neurophysiol* **107**, 2506-2516, doi:10.1152/jn.00719.2011 (2012).