Chapter 11: Learning to act

This chapter develops the theory of reinforcement learning for action selection. An agent can learn to optimize its actions through error-driven learning algorithms, implemented in striatal circuits that receive dopaminergic error signals. The theory is then generalized to consider the cognitive costs of action selection (quantified using information theory). A cost-sensitive learning algorithm can explain the origin of habits and a range of other apparently suboptimal behaviors. It is also compatible with data on the sensitivity of dopamine neuron activity to cognitive cost. Finally, we discuss how action selection algorithms balance exploration and exploitation during learning.

In the last chapter, we introduced simple algorithms for estimating value functions. We now show how these algorithms can be "put to work" in the service of action selection. Specifically, we study *policy optimization*—the search for reward-maximizing mappings from states to actions (or more precisely, distributions over actions). Policy optimization can be done efficiently by following gradients. We will also see how this leads to a neural implementation resembling what is seen in the basal ganglia. The second part of this chapter expands this picture to encompass the diverse ways in which animals learn to select actions.

1 *Policy optimization*

Recall that a policy defines a distribution over action a conditional on state s. We will use $V^\pi(s)$ to denote the value (expected discounted future return) of state s under policy π:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, \pi\right] \\ &= \sum_a \pi(a|s) \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a, \pi\right] \\ &= \sum_a \pi(a|s) Q^\pi(s,a), \end{aligned} \tag{1}$$

where r_t is the reward received at time t and γ is a discount factor. This definition is the same as the one given in the last chapter, only conditional on actions selected according to π. The term $Q^\pi(s,a)$ is the state-action value function, representing the value of taking action a in state s.

As we saw in the last chapter, the key to efficient reinforcement learning (RL) algorithms is the assumption of Markovian dynamics, which allowed us to decompose the value function recursively (the Bellman equation). We can do the same thing here by assuming that the transition dynamics are Markovian conditional on actions, with a transition distribution $T(s'|s,a)$. Equipped with this assumption, the environment is known as a *Markov decision process* (MDP).

The state transition distribution from the last chapter can be obtained by marginalizing over actions: $T^{\pi}(s'|s) = \sum_a \pi(a|s)T(s'|s,a)$.

We will make one additional assumption that is fairly innocuous but necessary for some of the results below. Typically, RL algorithms require for convergence that every state-action pair is sampled, which means that every state is accessible from every other state. This is known as *irreducibility*. Every irreducible MDP has a unique *stationary distribution*, $\mu^{\pi}(s) = \lim_{t\to\infty} p(s_t = s|s_0,\pi)$, where s_0 is the initial state. In other words, the stationary distribution defines the probability of visiting states after following π for long enough.

The policy optimization problem is to maximize the value, taking the expectation over the stationary distribution:

$$\pi^* = \operatorname*{argmax}_{\pi} \sum_s \mu^{\pi}(s)V^{\pi}(s). \tag{2}$$

Solving this problem efficiently generally requires gradient-based algorithms (see Chapter 9).

1.1 Policy gradient algorithms

In order to apply gradient-based algorithms, we need to assume that the policy is a differentiable function of parameters θ, so that we can write it as $\pi_\theta(a|s)$. We then define the objective function as:

$$\mathcal{J}(\theta) = \sum_s \mu^{\pi}(s)V^{\pi}(s). \tag{3}$$

To keep the notation light, we will sometimes leave the dependence of π on θ implicit.

To apply gradient descent, $\Delta\theta \propto \nabla_\theta \mathcal{J}(\theta)$, we need the gradient of the objective function with respect to the parameters, $\nabla_\theta \mathcal{J}(\theta)$. The *policy gradient theorem* (Sutton et al., 1999) provides a useful expression for this gradient:

$$\nabla_\theta \mathcal{J}(\theta) = \sum_s \mu^{\pi}(s) \sum_a Q^{\pi}(s,a)\nabla_\theta \pi_\theta(a|s). \tag{4}$$

There are many ways to compute an unbiased estimate of the policy gradient. The most straightforward way is to compute a stochastic gradient based on trajectories of states $\{s_t\}$ sampled from the transition distribution T^{π} (which will converge eventually to μ^{π}):

The expected difference between an unbiased estimate and the true value is 0.

$$\nabla_\theta \mathcal{J}(\theta) \approx \sum_a Q^{\pi}(s_t,a)\nabla_\theta \pi_\theta(a|s_t). \tag{5}$$

We can take this one step further and use the sampled actions $\{a_t\}$:

$$\begin{aligned} \nabla_\theta \mathcal{J}(\theta) &= \sum_a \pi_\theta(a|s_t) Q^{\pi}(s_t,a) \frac{\nabla_\theta \pi_\theta(a|s_t)}{\pi_\theta(a|s_t)} && (6) \\ &\approx Q^{\pi}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t). && (7) \end{aligned}$$

Dividing by the policy compensates for the fact that we are replacing a sum with an average.

We thus have a fully online algorithm that can operate on sampled state-action trajectories.

In practice, we typically don't have access to $Q^\pi(s,a)$, so we need to estimate it. We can do this using a state-action version of the TD algorithm we applied to state value estimation in the last chapter. However, this runs into several problems. One is that it will generally require function approximation over the state-action space, which is computationally costly and may require many samples to estimate accurately. Second, the policy gradient estimate has high variance due to the fact that state-action values may differ considerably across states even when this doesn't affect the policy.

To get some intuition for this, consider what happens when you're trying to decide what to buy for dinner at two grocery stores. These stores carry the same goods, but one store has higher prices (e.g., because it's located somewhere with greater inflation). The optimal policy is the same for both stores because the *relative* values are the same (you'd prefer the same foods in both cases), but the *absolute* values are different. In other words, there is an action-independent, state-dependent component to the state-action values, which contributes to the variance of the gradient. Ideally, we could remove this component and thereby reduce the variance, accelerating learning.

1.2 *Actor-critic algorithms*

A natural way to remove the state-dependent component is by subtracting the state values $V^\pi(s)$ from the state-action values. This produces what are called *advantages* (Baird, 1994):

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s). \tag{8}$$

Importantly, replacing $Q^\pi(s,a)$ with $A^\pi(s,a)$ reduces variance without distorting the policy gradient (it is still unbiased in expectation). The next step is to obtain an unbiased estimate of the advantages. This can be done with the TD errors themselves. Recall the definition of the TD error for state value learning described in the last chapter:

$$\delta = r + \gamma \hat{V}^\pi(s') - \hat{V}^\pi(s), \tag{9}$$

The only change here from the last chapter is that now the value function is policy-dependent.

where $\hat{V}^\pi(s)$ is an approximation of $V^\pi(s)$. When conditioning on a state-action pair, the expectation of the first two terms is the state-action value:

$$\mathbb{E}[r + \gamma \hat{V}^\pi(s')|s,a,\pi] = \hat{Q}^\pi(s,a), \tag{10}$$

an approximation of $Q^\pi(s,a)$. Thus, the expected TD error for a given state-action pair is equal to an approximation of the advantage:

$$\mathbb{E}[\delta|s,a,\pi] = \hat{A}^\pi(s,a). \tag{11}$$

This means we can use the TD error to define an unbiased policy gradient approximation:

$$\nabla_\theta \mathcal{J}(\theta) \approx \delta \nabla_\theta \log \pi_\theta(a|s). \tag{12}$$

This algorithm is known as an *actor-critic* method (Barto et al., 1983) because it relies on the interplay between an actor (the policy) and a critic (the value function estimator). TD errors, generated by the value function estimator, signal whether the current policy is producing higher or lower rewards than expected). In the next section, we will see how this interplay might be realized in the brain.

2 *Neurobiology of instrumental learning*

Motor control in the brain has a peculiar anatomical arrangement (Figure 1) where thalamic neurons controlling movement initiation (via connections to premotor cortex) are under tonic inhibition from the output nuclei of the basal ganglia, the globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr). Thus, in some sense movements are always "ready to go" upon disinhibition of the thalamus. This happens when GPi and SNr are themselves inhibited by upstream structures in the basal ganglia: a "direct" pathway from the dorsal striatum (caudate and putamen), and an "indirect" pathway from the dorsal striatum through the globus pallidus external segment (GPe). The direct pathway promotes action production (disinhibition of the thalamus), whereas the indirect pathway suppresses action production. This is why the direct pathway is sometimes referred to as a "Go" pathway and the indirect pathway as a "NoGo" pathway (Frank, 2005).

Although this description seems dizzyingly complicated, it's actually an oversimplification!

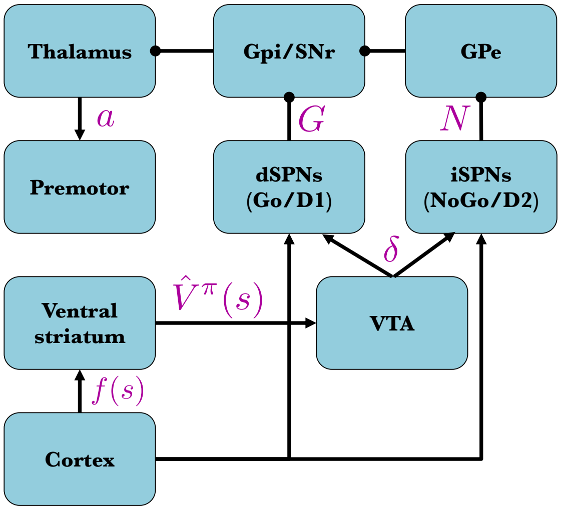

Figure 1: **Simplified diagram of the action selection circuit**. Excitatory connections are denoted by arrows; inhibitory connections are denoted by circles. VTA: ventral tegmental area (other labels are defined in the text).

Within the dorsal striatum, separate populations of medium spiny neurons project to the direct and indirect pathways. These are labeled the direct/indirect spiny projection neurons (dSPNs and iSPNs). These neurons receive the same cortical inputs, and therefore play a pivotal role in mapping state representations to action probabilities. Some of their functional differences arise from their distinct responses to dopamine inputs from the midbrain:

- The dSPNs express D1 dopamine receptors, whereas the iSPNs express D2 receptors. Dopamine has opposite effects on these neuron types, exciting dSPNs and inhibiting iSPNs (Surmeier et al., 2007).

- Because D2 receptors have a higher affinity for dopamine (Richfield et al., 1989), the inhibitory effects of dopamine predominate at low concentrations.

- Because D2 receptors saturate at relatively low concentrations compared to D1 receptors (Richfield et al., 1987), the excitatory effects of dopamine predominate at high concentrations.

- Dopamine signaling induces synaptic plasticity with opposite signs, promoting potentiation in dSPNs and depression in iSPNs (Calabresi et al., 2007; Shen et al., 2008). Plasticity at corticostriatal synapses follows a three-factor Hebbian rule (coincidence of presynaptic and postsynaptic firing with dopamine).

2.1 *Striatal policy parametrization*

Abstracting away from some of the anatomical and physiological details, we can synthesize most of these facts in a simple model, where dSPNs and iSPNs push in opposite directions:

The model described in this section is heavily influenced by (but not identical to) models described in prior work (Collins and Frank, 2014; Mikhael and Bogacz, 2016; Jaskir and Frank, 2023; Pinto and Uchida, 2025).

$$\pi(a = j|s) \propto \exp\left[\alpha^G G_j - \alpha^N N_j\right], \tag{13}$$

where G_j is the input to the dSPNs ("Go" neurons) tuned to action j, and N_j is the input to the iSPNs ("NoGo" neurons) tuned to action j. These inputs are modeled as linear combinations of cortical state features, $x_d = f_d(s)$, where $f_d(s)$ is the tuning function for cortical neuron d:

Note that these synapses are between cortex and dorsal striatum, distinct from the synapses between cortex and ventral striatum parametrizing the value function approximation, as described in the last chapter.

$$G_j = \sum_d \theta^G_{dj} x_d, \qquad N_j = \sum_d \theta^N_{dj} x_d. \tag{14}$$

The policy parameters can thus be interpreted as corticostriatal synaptic strengths.

The sensitivity parameters α^G and α^N reflect the modulatory influence of tonic dopamine ρ on dSPNs and iSPNs, respectively. The function relating ρ to sensitivity is based on the dose-occupancy functions for D1 and D2 receptors, combined with their postsynaptic effects, which we approximate as sigmoids (Figure 2):

$$\alpha^G = 1 + \tanh(\rho), \qquad \alpha^N = 1 - \tanh(\rho), \tag{15}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function that maps tonic dopamine levels to $[-1, 1]$. The specific mathematical assumptions here are somewhat ad hoc; what's important is that high levels of tonic dopamine amplify dSPNs and suppress iSPNs, whereas low levels suppress dSPNs and amplify iSPNs.

With this parametrization, we can write the policy gradient update as:

$$\Delta\theta^G_{dj} \propto \alpha^G \delta x_d y_j, \qquad \Delta\theta^N_{dj} \propto -\alpha^N \delta x_d y_j, \tag{16}$$

where

$$y_j = \mathbb{I}[a = j] - \pi_\theta(a = j|s) \tag{17}$$

can be interpreted as an *action prediction error*; y_j is positive when action j occurs unexpectedly, and is negative when action j is expected but fails to occur.

In order to interpret the updates in Eq. 16 as 3-factor Hebbian rules (presynaptic × postsynaptic × TD error), y_j must correspond to the postsynaptic (striatal) activity, and this must be the same for both dSPNs and iSPNs. Thus, dSPNs and iSPNs should be negatively correlated prior to a decision (since they push action selection in opposite directions), but should be positive correlated after action selection (to implement the policy gradient update in a biologically plausible fashion). This is precisely what was found in an analysis of SPN recordings (Lindsey et al., 2025). The same dataset (Markowitz et al., 2018) provides specific evidence supporting the action prediction error hypothesis: both cell types show higher activity following a low probability action compared to a high probability action.

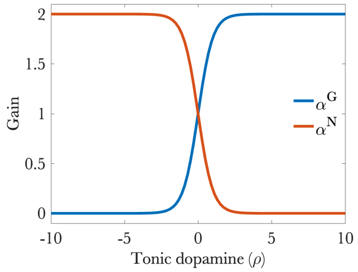

Figure 2: **Sensitivity for Go and NoGo components as a function of tonic dopamine**.

2.2 *The role of opponency*

The Go/NoGo model was originally motivated by clinical observations (Albin et al., 1989). "Hyperkinetic" movement disorders are characterized by excessively fast movements, leading to control failures. For example, people with chorea (derived from the Greek word for "dance") experience rapid, intrusive movements (Figure 3). This disorder often arises from Huntington's disease, which is associated with neurodegeneration in the striatum, particularly in the indirect pathway (Deng et al., 2004). D2 receptor antagonists, which should

Figure 3: **A child with chorea**. Reproduced from La Médecine Illustrée 1880.

have the effect of increasing activity in the indirect pathway (due to dopamine's inhibitory effect on iSPNs) are a standard treatment for hyperkinesia.

In contrast, "hypokinetic" movement disorders are characterized by excessively slow movements (bradykinesia), inability to initiate movements (akinesia), and rigidity of movement. Hypokinesia is a classic symptom of Parkinson's disease, which is associated with reduced striatal dopamine levels. Due to the higher affinity of D2 receptors, reduced dopamine levels leads to domination of action-suppressing iSPNs. L-DOPA, a standard treatment for Parkinson's, increases striatal dopamine and rebalances the relative activity of the direct and indirect pathways. In some cases, Parkinson's patients experience overdoses of dopamine leading to hyperkinesia, supporting the view that dopamine functions as a continuous knob that can shift the overall propensity for movement.

These clinical observations have been buttressed by studies in rodents which directly intervene on the two pathways. Stimulating dSPNs produces hyperkinetic symptoms, whereas stimulating iSPNs produces hypokinetic symptoms (Kravitz et al., 2010). Similar effects were observed in mice with genetic knockouts that selectively impaired one of the pathways (Bateup et al., 2010).

In summary, the direct and indirect pathways exert opposing influences on movement control. Disrupting the balance between the pathways can produce too little or too much movement. However, the opponency is doing more than controlling the overall level of movement—it's also controlling the impact of rewards and punishments on movement selection. A good illustration of this point comes from a study by Yttri and Dudman (2016), who stimulated dSPNs or iSPNs after movements that were either slower or faster than usual. They found that pairing dSPN stimulation with a particular movement speed reinforced that speed regardless of whether it was fast or slow, whereas pairing iSPN stimulation with a movement speed suppressed that speed, again regardless of whether it was fast or slow.

Going beyond movement speed, we can see the implications of opponency in action selection. For example, Kravitz et al. (2012) trained mice to optogenetically self-stimulate either iSPNs or dSPNs by pressing a trigger. Mice that received dSPN stimulation exhibited a tendency to repeatedly press the trigger, whereas mice that received iSPN stimulation exhibited a tendency to avoid pressing the trigger. Similar results have been obtained when optogenetic stimulation of dSPNs or iSPNs was paired with natural reward (Nonomura et al., 2018).

2.3 *The role of the critic*

In the last chapter, we presented evidence that the ventral striatum (nucleus accumbens) is responsible for encoding an approximation of the state value function, $\hat{V}$. The parameters of this function approximator are the synapses linking cortical inputs to neurons in the ventral striatum, updated by TD learning using the error signal δ conveyed by dopamine. The same TD error, we have argued here, is used to update the policy parameters (synapses linking cortical inputs to the dorsal striatum).

This architecture has several empirical implications. First, it implies a division of labor between ventral and dorsal striatum, where it is principally dorsal striatum that tracks action preferences. This is consistent with electrophysiology studies showing that action preference signals are prevalent in dorsal striatum (Samejima et al., 2005; Pasquereau et al., 2007; Lau and Glimcher, 2008) but are typically weak or absent in ventral striatum (Kim et al., 2009; Ito and Doya, 2015). Second, it implies that the ventral striatum, but not the dorsal striatum, should be active during classical conditioning, when value learning (but not policy updating) is engaged, whereas both regions should be active during instrumental conditioning (when both value and policy updating are engaged). This is consistent with evidence from human brain imaging (O'Doherty et al., 2004). Third, it implies that the error signal is uniform across ventral and dorsal striatum. This is consistent with recordings of dopamine neuron axons projecting to different parts of the striatum (Tsutsui-Kimura et al., 2020), though other work suggests more regional heterogeneity (van Elzelingen et al., 2022).

The mapping of the actor onto dorsal striatum and the critic onto ventral striatum was first proposed by Houk et al. (1995).

At the behavioral level, the actor-critic model implies a form of state dependence. Let's imagine a choice between two actions, a_1 and a_2, which have been rewarded the same number of times, but in different states; action a_1 was previously chosen in a "rich" state (high reward availability), whereas action a_2 was previously chosen in a "poor" state (low reward availability). Because the rich state has a higher value than the poor state, the TD error will be smaller in the rich state. Thus, action a_2 will acquire a stronger preference than action a_1, consistent with studies in several species (Pompilio and Kacelnik, 2005; Pompilio et al., 2006; Aw et al., 2009; Palminteri et al., 2015).

Another behavioral implication is learning about *unchosen* (counterfactual) actions: all policy parameters are updated after receiving feedback, even those parameters corresponding to unchosen actions. The updates for these parameters are sign-reversed, because they drive the policy away from the chosen action, consistent with experimental studies in humans (Biderman and Shohamy, 2021; Biderman et al., 2023; Ben-Artzi et al., 2023). Specifically, getting rewarded for choosing a_1 over a_2 will make it less likely that a_2 will be chosen later, even if a_1 is not in the set of available actions. Action a_2 is more likely to be chosen later if choosing a_1 was instead unrewarded.

Biderman and Shohamy (2021) name this effect the *inverse decision bias*.

2.4 *The role of tonic dopamine*

In the Go/NoGo model, tonic dopamine influences both the relative activation of the Go and NoGo components, as well as the effective learning rates of each component. The relative activation controls risk preferences, as discussed in the next section. The effective learning rates control how much is learned from positive outcomes (driving the Go component) vs. negative outcomes (driving the NoGo component). When tonic dopamine is high, the effective learning rate for the Go component is large, and the effective learning rate for the NoGo component is close to 0. When tonic dopamine is low, the effective learning rate for the NoGo component is large, and the effective learning rate for the Go component is close to 0.

One consequence of this pattern is that high tonic dopamine levels should facilitate learning from positive outcomes, whereas low tonic dopamine levels should facilitate learning from negative outcomes. In support of this prediction, Frank et al. (2004) showed that Parkinson's patients off dopamine medication are better at learning from negative outcomes than from positive outcomes, whereas patients on medication show the opposite pattern. Along the same lines, increasing tonic dopamine using the stimulant methylphenidate induces greater sensitivity to rewards and reduced sensitivity to cognitive effort costs (Westbrook et al., 2020).

Willingness to exert effort may also be controlled through the relative activation of the Go and NoGo components, independently of learning effects. Under high tonic dopamine, the Go component is amplified and the NoGo component is suppressed, such that the benefits of effort are weighed more heavily than the costs. This prediction is broadly consistent with the literature on the effects of dopamine depletion, which typically reduce willingness to exert cognitive or physical effort. For example, rodents trained on "fixed ratio" tasks can earn more reward (food pellets) by pressing a response lever more frequently. The ratio schedule determines how many presses are required to produce a pellet delivery. Because lever pressing is effortful, rodents will only press at a higher rate if motivated by a higher ratio (i.e., more pressing is necessary to get the same amount of reward). Dopamine depletion blunts this motivational effect: rodents not only press overall less frequently, but also show reduced sensitivity to the ratio schedule (Aberman and Salamone, 1999). Similarly, dopamine-depleted rodents are less likely to climb over a barrier to obtain a more desirable food reward (Salamone et al., 1994). However, dopamine depletion does not affect their propensity to approach the same food in the absence of a barrier, indicating that this effect is due to motivation rather than anhedonia—i.e., they like the reward just as much with or without dopamine, but the absence of dopamine reduces their willingness to work for the reward.

One way to understand tonic dopamine's motivational effects is through the lens of average reward, an idea we encountered at the end of Chapter 6 in the context of attention. Niv et al. (2007) posited that tonic dopamine tracks average reward, which in turn defines the opportunity cost of effort (i.e., how much reward is foregone if effort is not exerted). According to this hypothesis, dopamine depletion reduces motivation by altering the signal used to compute the benefits of effort exertion. This is consistent with the finding that dopamine fluctuations on the timescale of minutes covary with both reward rate and response vigor (Hamid et al., 2016).

In order for the average reward hypothesis of tonic dopamine to be consistent with the TD error hypothesis of phasic dopamine, it is necessary for average rewards to reflect a slow averaging of TD errors (Gershman et al., 2024). This does not arise automatically for discounted value functions, but it does for a slightly different formulation (Mahadevan, 1996), where the (undiscounted) value function is defined as the state-dependent average reward relative to the average reward:

The discounted and average reward formulations are closely related, approaching one another as $\gamma \rightarrow 1$ (Kakade, 2001).

$$V^{\pi}(s) = \lim_{H \rightarrow \infty} \frac{1}{H} \sum_{t=1}^{H} \mathbb{E}[r_t - \bar{r} | s, \pi], \tag{18}$$

Where $\bar{r}$ is the average reward. The average reward reference point is needed to ensure that the infinite sum doesn't diverge. The corresponding TD error is defined accordingly:

$$\delta = r - \bar{r} + V(s') - V(s). \tag{19}$$

Thus, the TD error is referenced to the average reward. Importantly, it can be shown that temporally averaging these TD errors into a tonic signal yields an estimate of average reward (Wan et al., 2021).

Linking the TD error in Eq. 19 to the hypothesis that tonic dopamine encodes the average reward ($\rho \approx \bar{r}$) is consistent with the antagonistic effect of tonic dopamine on phasic release via the action of autoreceptors (Grace, 1991; Benoit-Marand et al., 2001). One reflection of this antagonism is the gradual decrease of phasic responses as reward

Autoreceptors are receptors that bind to the release products of a cell, typically inhibiting subsequent release (a form of negative feedback control).

rate increases (Kilpatrick et al., 2000), consistent with average-reward TD modeling (Daw and Touretzky, 2002).

2.5 *Risk sensitivity*

Dopaminergic modulation of the direct and indirect pathways provides a mechanism for parametrizing risk sensitivity. Consider a choice between a risky option that delivers reward R with probability P (otherwise 0) and a safe option that always delivers reward $S < R$. The *certainty equivalent* is the value of S that would be required to make an agent indifferent between the risky and safe options. The *risk premium* is the difference between the expected payoff for the risky option (RP in this case) and the certainty equivalent; it quantifies how much an agent is willing to pay to avoid the risk—a measure of their *risk aversion*. While humans are typically risk averse for positive outcomes, they tend to be risk seeking for negative outcomes, preferring risky over safe options with the same expected value (Kahneman and Tversky, 1979).

To understand these patterns of risk sensitivity through the lens of the Go/NoGo model, we will reformulate the model (following Mikhael and Bogacz, 2016) to make its risk preferences more transparent. We start by rewriting the net drive $D_j = \alpha^G G_j - \alpha^N N_j$ for option j as follows:

$$D_j \propto (\alpha^G + \alpha^N)(G_j - N_j) + (\alpha^G - \alpha^N)(G_j + N_j). \qquad (20)$$

Asymptotically, $G_j - N_j \propto \mu_j$, the expected reward for option j, and $G_j + N_j \propto \sigma_j$, the reward standard deviation. Plugging these into Eq. 20 gives:

$$D_j \propto \mu_j + \beta\sigma_j, \qquad \beta = \frac{\alpha^G - \alpha^N}{\alpha^G + \alpha^N}. \qquad (21)$$

Thus, the asymptotic output of the direct and indirect pathways can be viewed as a linear combination of mean and standard deviation components. When $\alpha^G - \alpha^N = 0$, the agent is risk-neutral ($\lambda = 0$): the action probability depends only on the mean rewards. When $\alpha^G < \alpha^N$, the agent is risk-averse ($\lambda < 0$): the action probability decreases with the standard deviation. Finally, when $\alpha^G > \alpha^N$, the agent is risk-seeking ($\lambda > 0$): the action probability increases with the standard deviation.

Recall that tonic dopamine is hypothesized to increase α^G and decrease α^N. This means that λ will increase with tonic dopamine, producing risk aversion at low levels and risk seeking at high levels. This is consistent with numerous studies. Unmedicated Parkinson's patients (low tonic dopamine) are relatively more risk-averse than healthy controls, and this difference is eliminated by dopaminergic medication (Cherkasova et al., 2019). Medication can even cause pathological gambling, which is reduced after cessation of medication (Dodd et al., 2005). In healthy humans, boosting dopamine with L-DOPA increases risk-seeking (Rutledge et al., 2015; Rigoli et al., 2016a). Pharmacologically blocking D2 receptors (effectively reducing α^N) increases risk seeking (Burke et al., 2018), whereas activating D2 receptors decreases risk seeking (Simon et al., 2011). Similarly, optogenetically stimulating iSPNs can convert risk seeking to risk aversion (Zalocusky et al., 2016).

Under the hypothesis that tonic dopamine tracks reward rate, we can also make the prediction that risk seeking will increase with reward rate. Evidence for this prediction comes from several studies. Gilby and Wrangham (2007) found that risk-seeking in wild chimpanzees (operationalized as engaging in risky hunting rather than safe foraging) increases during periods of higher diet quality (greater availability of ripe fruit). Humans likewise take more risks when reward rate is high: people shift from risk aversion toward risk seeking immediately following a meal (Symmonds et al., 2010), a shift from low to high reward context (Rigoli et al., 2016b), and even after a single prior gain (Thaler and Johnson, 1990) or incidental positive outcome (e.g., a win by the local sports team; Otto et al., 2016).

2.6 *Exploration*

A fundamental problem in RL is the *exploration-exploitation dilemma*: to maximize long-term reward, an agent needs to balance exploring (gathering information about possibly low-reward actions) and exploiting (choosing actions that are believed to be best based on current estimates, which might be wrong). The optimal solution to this problem is intractable, but many heuristics have been invented, some with theoretical guarantees. One influential heuristic (Auer, 2002) is to choose options based on the sum of a mean reward estimate $\hat{\mu}_j$ and an uncertainty bonus $\beta\sigma_j$. This looks a lot like Eq. 21, and it is! However, the uncertainty bonus heuristic requires that $\beta \geq 0$, whereas Eq. 21 allows $\beta < 0$ (risk aversion) if $\alpha^N > \alpha^G$. Moreover, this situation will tend to happen early during learning, when reward rate (and hence tonic dopamine) is low.

Behavioral (Gershman, 2018, 2019) and neuroimaging (Tomov et al., 2020) studies provide evidence that humans use an uncertainty bonus to guide exploration.

To address this puzzle, we can appeal to another aspect of dopamine—novelty responses. It has been argued that the elevated response of dopamine neurons to novelty poses a challenge to RL theories of dopamine (Horvitz, 2000; Kutlu et al., 2021), but it might actually be part of the solution to the exploration-exploitation dilemma faced by all RL algorithms (Kakade and Dayan, 2002; Wang et al., 2024).

Generally speaking, novelty is a proxy for uncertainty, because agents will be most uncertain about the value of novel stimuli. As the agent gets more experience with the stimulus, uncertainty decreases along with novelty. Thus, transiently boosting tonic dopamine (ρ) to novel stimuli could be a way to implement an uncertainty bonus. This would have the effect of differentially activating dSPNs more than iSPNs (i.e., $\alpha^G > \alpha^N$), thereby yielding $\beta > 0$ (risk seeking). Importantly, the exploratory boost would diminish over the course training, so that agents could eventually converge on their asymptotic risk preference.

See the related discussion about uncertainty and latent inhibition in the last chapter.

Dopamine can also facilitate exploitation by decreasing decision or sensory noise (Gershman and Tzovaras, 2018; Mikhael et al., 2021; Chen et al., 2024).

Consistent with this idea, pharmacologically elevating dopamine (by inhibiting dopamine reuptake) increases novelty seeking in monkeys (Costa et al., 2014). The specific importance of D1 receptors is supported by the finding that antagonizing D1 receptors reduces novelty seeking (Peters et al., 2007). At the behavioral level, average reward has been shown to increase novelty seeking (Gershman and Niv, 2015), consistent with the average reward model of dopamine. The same relationship is seen in a naturalistic setting: people are more likely to try new restaurants if they live in areas where the average restaurant quality is high (Schulz et al., 2019). Furthermore, this novelty preference is accentuated when the restaurant quality has high variance, consistent with the use of an uncertainty bonus.

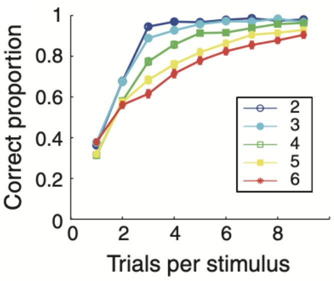

Figure 4: **Performance degrades with set size**. On each trial, subjects had to choose the correct action for a specific stimulus (indicating the state). The number of distinct stimuli in a block is the set size. Each curve shows the proportion of correct actions as a function of trial in a learning block for a given set size. Reproduced from Collins and Frank (2012).

3 *Policy compression*

When the number of states is increased, human performance starts to decline (Figure 4). Evidently there is a representational or memory constraint on RL, which is not built into the model developed above. One way to understand this constraint is that it reflects a limit on the state encoding function $f(s)$; if this produces representations (x) that overlap across states, then there can potentially be confusion between states. Alternatively, the mapping from representations to action probabilities may be constrained (e.g., the policy weights θ can't get too large). More generally, we can quantify how state-dependent the policy is using the mutual information between states and actions, $\mathcal{I}[s;a]$. We will refer to this quantity as the *policy complexity* to capture the intuition that policies are more complex when they are more sensitive to variations in the state (Gershman, 2020; Lai and Gershman, 2021).

Set size effects like the one shown in Figure 4 can be explained by imposing an upper bound (capacity limit) $\mathcal{C}$ on policy complexity. We can then study the achievable average reward for a given capacity limit. This yields a *reward-complexity frontier* (Figure 5). Each policy maps to a single point in the reward-complexity plane; policies below

The concepts applied here are derived from *rate-distortion theory*, which we will revisit in Chapter 13.

the frontier are suboptimal, and policies above the frontier are unachievable. The capacity limit corresponds to a vertical slice through the reward-complexity plane. The point where it intersects the frontier corresponds to the set of optimal policies for an agent with that capacity limit.

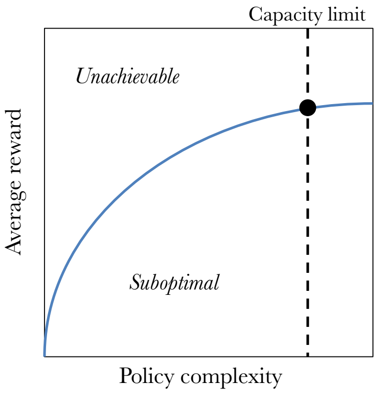

Figure 5: **The reward-complexity plane**. The curve show the reward-complexity frontier, separating unachievable from suboptimal policies. The circle shows the optimal achievable average reward for a given capacity limit (upper bound on policy complexity).

We can write the capacity-limited reward optimization problem as a Lagrangian (Tishby and Polani, 2010; Still and Precup, 2012):

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \rho(\pi) - \lambda c(\pi), \tag{22}$$

$$\lambda = \frac{\partial \rho(\pi)}{\partial c(\pi)}, \tag{23}$$

where we have expressed the average reward $\rho(\pi)$ as a function of the policy, and $c(\pi)$ is the complexity of π. The parameter $\lambda \geq 0$ is a Lagrange multiplier that monotonically decreases with the capacity limit $\mathcal{C}$ (more precisely, it's the slope of the reward-complexity frontier at the point where it intersects the capacity limit). In the limit $\mathcal{C} \to \infty$ (no bound on policy complexity), $\lambda \to 0$ and we recover average reward optimality.

The optimal solution to Eq. 22 can be written explicitly:

This analysis is closely related to the analysis of approximate inference in Chapter 4.

$$\pi^*(a|s) \propto \exp\left[Q^\pi(s,a) + \lambda \log p^*(a)\right], \tag{24}$$

$$p^*(a) = \sum_s \pi^*(a|s) p(s). \tag{25}$$

Eq. 24 is remarkable for several reasons. One is that it takes the form of a softmax policy, which is usually imposed as an *ad hoc* parametrization to produce exploratory behavior (Sutton and Barto, 2018); here, it is derived from the capacity-limited optimization problem. Importantly, this policy produces stochasticity even asymptotically and under perfect knowledge of rewards; it reflects cognitive resource constraints rather than exploration (although it can induce useful exploration as a side effect). Another remarkable aspect of the policy is that it introduces a response bias $p^*(a)$ into the softmax, reflecting frequently chosen actions across all states. This bias only appears when capacity is limited ($\lambda > 0$). Human behavioral studies have shown that such a bias exists, and that the bias increases with set size (Lai and Gershman, 2024), consistent with the idea that the bias arises from a limited resource that is shared across states. For a similar reason, choice stochasticity increases with set size (Lai and Gershman, 2021).

The specific functional form of perseverative bias in Eq. 24 was supported by model-based analyses in Gershman (2020).

Because the policy is continuously changing during learning, the bias needs to be incrementally updated. This predicts a form of perseveration or stickiness, where agents continue trying actions they chose frequently in the past, independently of the reward history. Perseveration is a well-established phenomenon across many different tasks (e.g., Collier et al., 1952; Lau and Glimcher, 2005; Worthy et al., 2013)—so much so that Thorndike (1911) elevated it to a "Law of Exercise." The reward-complexity analysis allows us to test whether perseverative biases are "optimal" in the sense that the policies lie close to the reward-complexity frontier. Several studies have shown that attested policies are in fact quite close to the frontier (Gershman, 2020; Lai and Gershman, 2024; Gershman and Lak, 2025), though there tends to be a fall-off for low-complexity policies, which might arises from differences in learning (e.g., learning rates, as discussed in Gershman and Lai, 2021).

The fall-off is particularly pronounced in patients with schizophrenia (Gershman and Lai, 2021).

We can derive a learning algorithm for the capacity-limited optimization problem by noting that it is equivalent to:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[r - \lambda \log \frac{\pi(a|s)}{p^*(a)}\right]. \tag{26}$$

This means that a standard policy gradient algorithm can be used to find the optimal policy, simply by adding a complexity penalty (how much the state-dependent policy deviates from the action bias) to the rewards. This predicts that phasic dopamine signals encoding TD errors should be suppressed by policy complexity, as observed empirically (Gershman and Lak, 2025).

4 *Conclusion*

The RL machinery developed in the last chapter was put to work in this chapter for policy optimization. We showed how a biologically plausible policy parametrization, based on opponency in the direct and indirect pathways of the basal ganglia, could be used to learn optimal actions. Tonic dopamine played an important role in this architecture, governing both exploration during learning and asymptotic risk sensitivity. Finally, we showed how augmenting the reward function with a complexity penalty enabled capacity-limited policy optimization, which naturally explained behavioral stochasticity and perseveration, as well as the sensitivity of dopamine neuron activity to policy complexity.

Our treatment of RL so far has been somewhat narrow; the only objective is to predict and maximize reward. However, there is evidence that the brain is capable of learning richer representations of the world, and to use these representations in the service of flexible, goal-directed behavior. We will see in the next chapter how the same RL machinery can be generalized to support learning these richer representations.

Study questions

1. Why might evolution have favored an actor-critic division in the brain?
2. How can we reconcile hypothetical phasic dopamine encoding of prediction errors with tonic dopamine encoding of average reward?
3. Novelty responses in dopamine neurons may function as "uncertainty bonuses" for exploration. How does this mechanism help resolve the exploration-exploitation dilemma, and how might it fail under pathological conditions?

References

Aberman, J. and Salamone, J. D. (1999). Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience*, 92:545–552.

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

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Aw, J., Holbrook, R., de Perera, T. B., and Kacelnik, A. (2009). State-dependent valuation learning in fish: Banded tetras prefer stimuli associated with greater past deprivation. *Behavioural Processes*, 81:333–336.

Baird, L. C. (1994). Reinforcement learning in continuous time: advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks*, volume 4, pages 2448–2453. IEEE.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846.

Bateup, H. S., Santini, E., Shen, W., Birnbaum, S., Valjent, E., Surmeier, D. J., Fisone, G., Nestler, E. J., and Greengard, P. (2010). Distinct subclasses of medium spiny neurons differentially regulate striatal motor behaviors. *Proceedings of the National Academy of Sciences*, 107:14845–14850.

Ben-Artzi, I., Kessler, Y., Nicenboim, B., and Shahar, N. (2023). Computational mechanisms underlying latent value updating of unchosen actions. *Science Advances*, 9:eadi2704.

Benoit-Marand, M., Borrelli, E., and Gonon, F. (2001). Inhibition of dopamine release via presynaptic D2 receptors: time course and functional characteristics in vivo. *Journal of Neuroscience*, 21:9134–9141.

Biderman, N., Gershman, S., and Shohamy, D. (2023). The role of memory in counterfactual valuation. *Journal of Experimental psychology. General*, 152:1754–1767.

Biderman, N. and Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, 12:4648.

Burke, C. J., Soutschek, A., Weber, S., Raja Beharelle, A., Fehr, E., Haker, H., and Tobler, P. N. (2018). Dopamine receptor-specific contributions to the computation of value. *Neuropsychopharmacology*, 43:1415–1424.

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

Chen, C. S., Mueller, D., Knep, E., Ebitz, R. B., and Grissom, N. M. (2024). Dopamine and norepinephrine differentially mediate the exploration-exploitation tradeoff. *Journal of Neuroscience*, 44.

Cherkasova, M. V., Corrow, J. C., Taylor, A., Yeung, S. C., Stubbs, J. L., McKeown, M. J., Appel-Cresswell, S., Stoessl, A. J., and Barton, J. J. (2019). Dopamine replacement remediates risk aversion in Parkinson's disease in a value-independent manner. *Parkinsonism & Related Disorders*, 66:189–194.

Collier, G. H., Cotton, J. W., et al. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, 44:273–282.

Collins, A. G. and Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35:1024–1035.

Collins, A. G. and Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, 121:337–366.

Costa, V., Tran, V., Turchi, J., and Averbeck, B. (2014). Dopamine modulates novelty seeking behavior during decision making. *Behavioral Neuroscience*, 128:556–566.

Daw, N. D. and Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, 14:2567–2583.

Deng, Y., Albin, R., Penney, J., Young, A., Anderson, K., and Reiner, A. (2004). Differential loss of striatal projection systems in Huntington's disease: a quantitative immunohistochemical study. *Journal of Chemical Neuroanatomy*, 27:143–164.

Dodd, M. L., Klos, K. J., Bower, J. H., Geda, Y. E., Josephs, K. A., and Ahlskog, J. E. (2005). Pathological gambling caused by drugs used to treat Parkinson disease. *Archives of Neurology*, 62:1377–1381.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17:51–72.

Frank, M. J., Seeberger, L. C., and O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, 306:1940–1943.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, 6:277–286.

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394.

Gershman, S. J., Assad, J. A., Datta, S. R., Linderman, S. W., Sabatini, B. L., Uchida, N., and Wilbrecht, L. (2024). Explaining dopamine through prediction errors and beyond. *Nature Neuroscience*, 27:1645–1655.

Gershman, S. J. and Lai, L. (2021). The reward-complexity trade-off in schizophrenia. *Computational Psychiatry*, 5:38.

Gershman, S. J. and Lak, A. (2025). Policy complexity suppresses dopamine responses. *Journal of Neuroscience*, 45.

Gershman, S. J. and Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, 7:391–415.

Gershman, S. J. and Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, 120:97–104.

Gilby, I. C. and Wrangham, R. W. (2007). Risk-prone hunting by chimpanzees (Pan troglodytes schweinfurthii) increases during periods of high diet quality. *Behavioral Ecology and Sociobiology*, 61:1771–1779.

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

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

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, 96:651–656.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of Information Processing in the Basal Ganglia*. MIT Press.

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

Jaskir, A. and Frank, M. J. (2023). On the normative advantages of dopamine and striatal opponency for learning and choice. *Elife*, 12:e85107.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–292.

Kakade, S. (2001). Optimizing average reward using discounted rewards. In *International Conference on Computational Learning Theory*, pages 605–615. Springer.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, 15:549–559.

Kilpatrick, M., Rooney, M., Michael, D., and Wightman, R. (2000). Extracellular dopamine dynamics in rat caudate–putamen during experimenter-delivered and intracranial self-stimulation. *Neuroscience*, 96:697–706.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *Journal of Neuroscience*, 29:14701–14712.

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

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

Kutlu, M. G., Zachry, J. E., Melugin, P. R., Cajigas, S. A., Chevee, M. F., Kelly, S. J., Kutlu, B., Tian, L., Siciliano, C. A., and Calipari, E. S. (2021). Dopamine release in the nucleus accumbens core signals perceived saliency. *Current Biology*, 31:4748–4761.

Lai, L. and Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of Learning and Motivation*, volume 74, pages 195–232. Elsevier.

Lai, L. and Gershman, S. J. (2024). Human decision making balances reward maximization and policy compression. *PLOS Computational Biology*, 20:e1012057.

Lau, B. and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84:555–579.

Lau, B. and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, 58:451–463.

Lindsey, J. W., Markowitz, J., Gillis, W. F., Datta, S. R., and Litwin-Kumar, A. (2025). Dynamics of striatal action selection and reinforcement learning. *eLife*, 13:RP101747.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22:159–195.

Markowitz, J. E., Gillis, W. F., Beron, C. C., Neufeld, S. Q., Robertson, K., Bhagat, N. D., Peterson, R. E., Peterson, E., Hyun, M., Linderman, S. W., et al. (2018). The striatum organizes 3D behavior via moment-to-moment action selection. *Cell*, 174:44–58.

Mikhael, J. G. and Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS Computational Biology*, 12:e1005062.

Mikhael, J. G., Lai, L., and Gershman, S. J. (2021). Rational inattention and tonic dopamine. *PLoS Computational Biology*, 17:e1008659.

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

Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., Watanabe, M., Yamanaka, K., Enomoto, K., Chiken, S., et al. (2018). Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. *Neuron*, 99:1302–1314.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, 304:452–454.

Otto, A. R., Fleming, S. M., and Glimcher, P. W. (2016). Unexpected but incidental positive outcomes predict real-world gambling. *Psychological Science*, 27:299–311.

Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, 6:8096.

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C. E., and Boraud, T. (2007). Shaping of motor responses by incentive values through the basal ganglia. *Journal of Neuroscience*, 27:1176–1183.

Peters, J. R., Vallie, B., Difronzo, M., and Donaldson, S. T. (2007). Role of dopamine D1 receptors in novelty seeking in adult female Long-Evans rats. *Brain Research Bulletin*, 74:232–236.

Pinto, S. R. and Uchida, N. (2025). Tonic dopamine and biases in value learning linked through a biologically inspired reinforcement learning model. *Nature Communications*, 16:7529.

Pompilio, L. and Kacelnik, A. (2005). State-dependent learning and suboptimal choice: when starlings prefer long over short delays to food. *Animal Behaviour*, 70:571–578.

Pompilio, L., Kacelnik, A., and Behmer, S. T. (2006). State-dependent learned valuation drives choice in an invertebrate. *Science*, 311:1613–1615.

Richfield, E. K., Penney, J. B., and Young, A. B. (1989). Anatomical and affinity state comparisons between dopamine D1 and D2 receptors in the rat central nervous system. *Neuroscience*, 30:767–777.

Richfield, E. K., Young, A. B., and Penney, J. B. (1987). Comparative distribution of dopamine D-1 and D-2 receptors in the basal ganglia of turtles, pigeons, rats, cats, and monkeys. *Journal of Comparative Neurology*, 262:446–463.

Rigoli, F., Rutledge, R. B., Chew, B., Ousdal, O. T., Dayan, P., and Dolan, R. J. (2016a). Dopamine increases a value-independent gambling propensity. *Neuropsychopharmacology*, 41:2658–2667.

Rigoli, F., Rutledge, R. B., Dayan, P., and Dolan, R. J. (2016b). The influence of contextual reward statistics on risk preference. *NeuroImage*, 128:74–84.

Rutledge, R. B., Skandali, N., Dayan, P., and Dolan, R. J. (2015). Dopaminergic modulation of decision making and subjective well-being. *Journal of Neuroscience*, 35:9811–9822.

Salamone, J. D., Cousins, M. S., and Bucher, S. (1994). Anhedonia or anergia? effects of haloperidol and nucleus accumbens dopamine depletion on instrumental response selection in a t-maze cost/benefit procedure. *Behavioural Brain Research*, 65:221–229.

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

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., and Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 116:13903–13908.

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

Simon, N. W., Montgomery, K. S., Beas, B. S., Mitchell, M. R., LaSarge, C. L., Mendez, I. A., Banuelos, C., Vokes, C. M., Taylor, A. B., Haberman, R. P., et al. (2011). Dopaminergic modulation of risky decision-making. *Journal of Neuroscience*, 31:17460–17470.

Still, S. and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131:139–148.

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12.

Symmonds, M., Emmanuel, J. J., Drew, M. E., Batterham, R. L., and Dolan, R. J. (2010). Metabolic state alters economic decision making under risk in humans. *PloS One*, 5:e11090.

Thaler, R. H. and Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science*, 36:643–660.

Thorndike, E. (1911). *Animal Intelligence*. The Macmillan Company.

Tishby, N. and Polani, D. (2010). Information theory of decisions and actions. In *Perception-action cycle: Models, architectures, and hardware*, pages 601–636. Springer.

Tomov, M. S., Truong, V. Q., Hundia, R. A., and Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, 11:2371.

Tsutsui-Kimura, I., Matsumoto, H., Akiti, K., Yamada, M. M., Uchida, N., and Watabe-Uchida, M. (2020). Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. *Elife*, 9:e62390.

van Elzelingen, W., Goedhoop, J., Warnaar, P., Denys, D., Arbab, T., and Willuhn, I. (2022). A unidirectional but not uniform striatal landscape of dopamine signaling for motivational stimuli. *Proceedings of the National Academy of Sciences*, 119:e2117270119.

Wan, Y., Naik, A., and Sutton, R. S. (2021). Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR.

Wang, Y., Lak, A., Manohar, S. G., and Bogacz, R. (2024). Dopamine encoding of novelty facilitates efficient uncertainty-driven exploration. *PLOS Computational Biology*, 20:e1011516.

Westbrook, A., Van Den Bosch, R., Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., and Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, 367:1362–1366.

Worthy, D. A., Pang, B., and Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, 4:640.

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

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