On the normative advantages of dopamine and striatal opponency for learning and choice

Alana Jaskir[1*] **and Michael J Frank**[1*]

***For correspondence:**
alana_jaskir@brown.edu (AJ);
michael_frank@brown.edu (MJF)

[1]Department of Cognitive, Linguistic and Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, United States

Abstract Much research focuses on how the basal ganglia (BG) and dopamine (DA) contribute to reward-driven behavior. But BG circuitry is notoriously complex, with two opponent pathways interacting via several disinhibitory mechanisms, which are in turn modulated by DA. Building on earlier models, we propose a new model, OpAL*, to assess the normative advantages of such circuitry in cost-benefit decision making. OpAL* dynamically modulates DA as a function of learned reward statistics, differentially amplifying the striatal pathway most specialized for the environment. OpAL* exhibits robust advantages over traditional and alternative BG models across a range of environments, particularly those with sparse reward. These advantages depend on opponent and nonlinear Hebbian plasticity mechanisms previously thought to be pathological. Finally, OpAL* captures patterns of risky choice arising from manipulations of DA and environmental richness across species, suggesting that such choice patterns result from a normative biological mechanism.

Everybody wants the most they can possibly get
For the least they can possibly do
– Todd Snider, *"Easy Money"*

Introduction

Everyday choices involve integrating and comparing the subjective benefits and costs of potential actions. Moreover, the degree to which one prioritizes costs or benefits may vary between and even within individuals. For example, one may typically use food preference to guide their choice of restaurant, but be more likely to minimize costs (e.g., speed, distance, price) when only low quality options are available (only fast-food restaurants are open). In this paper, we evaluate the computational advantages of such context-dependent choice strategies and how they may arise from biological properties within the basal ganglia (BG) and dopamine (DA) system. We find that biological properties within this system – specifically, the presence of opponent striatal pathways, nonlinear Hebbian plasticity, and dynamic changes in dopamine as a function of reward history – confer decision making advantages relative to canonical reinforcement learning models lacking these properties.

In neural network models of such circuitry, the cortex "proposes" candidate actions available for consideration, and the BG facilitates those that are most likely to maximize reward and minimize cost (***Frank, 2005***; ***Ratcliff and Frank, 2012***; ***Franklin and Frank, 2015***; ***Gurney et al., 2015***; ***Dunovan and Verstynen, 2016***). These models are based on the BG architecture in which striatal

medium spiny neurons (MSNs) are subdivided into two major populations that respond in opponent ways to DA (due to differential expression of D1 and D2 receptors; ***Gerfen*** (***1992***)). Phasic DA signals convey reward prediction errors (***Montague et al., 1996***; ***Schultz et al., 1997***), amplifying both activity and synaptic learning in D1 neurons, thereby promoting action selection based on reward. Conversely, when DA levels drop, activity is amplified in D2 neurons, promoting learning and choice that minimizes disappointment (***Frank, 2005***; ***Iino et al., 2020***).

Empirically, the BG and DA have been strongly implicated in such motivated action selection and reinforcement learning across species. For example, in perceptual decisions, striatal D1 and D2 neurons combine information about veridical perceptual data with internal preferences based on potential reward, causally influencing choice toward the more rewarding options (***Doi et al., 2020***; ***Bolkan et al., 2022***). Further, striatal DA manipulations influence reinforcement learning (***Yttri and Dudman, 2016***; ***Frank et al., 2004***; ***Pessiglione et al., 2006a***), motivational vigor (***Niv et al., 2007***; ***Beeler et al., 2012***; ***Hamid et al., 2015***), cost-benefit decisions about physical effort (***Salamone et al., 2018***) and risky decision making. Indeed, as striatal DA levels rise, humans and animals are more likely to select riskier options that offer greater potential payout than those with certain but smaller rewards (***St Onge and Floresco, 2009***; ***Zalocusky et al., 2016***; ***Rutledge et al., 2015***), an effect that has been causally linked to striatal D2 receptor-containing subpopulations (***Zalocusky et al., 2016***).

However, for the large part, this literature has focused on the findings *that* DA has opponent effects on D1 and D2 populations and behavioral patterns, and not what the computational advantage of this scheme might be (i.e., *why*). For example, the Opponent Actor Learning (OpAL) model (***Collins and Frank, 2014***) summarizes the core functionality of the BG neural network models in algorithmic form, capturing a wide variety of findings of DA and D1 vs D2 manipulations across species (for review, ***Collins and Frank*** (***2014***); ***Maia and Frank*** (***2017***)). Two distinguishing features of OpAL (and its neural network inspiration), compared to more traditional RL models, are that (i) it relies on opponent D1/D2 representations rather than a single expected reward value for each action and (ii) learning in such populations is acquired through nonlinear dynamics, mimicking three-factor hebbian plasticity rules. This nonlinearity causes the two populations to evolve to specialize in discriminating between options of high or low reward value, respectively ***Collins and Frank*** (***2014***). It is also needed to explain pathological conditions such as learned Parkinsonism, whereby low DA states induce hyperexcitability in D2 MSNs, driving aberrant plasticity and in turn, progression of symptoms (***Wiecki et al., 2009***; ***Beeler et al., 2012***).

But why would the brain develop this nonlinear opponent mechanism for action selection and learning, and how could (healthy) DA levels be adapted to capitalize on it? A clue to this question lies in the observation that standard (non-biological) RL models typically perform worse at selecting the optimal action in "lean environments" with sparse rewards than they do in "rich environments" with plentiful rewards (***Collins and Frank, 2014***). This asymmetry results from a difference in exploration / exploitation tradeoffs across such environments. In rich environments, an agent can benefit from overall higher levels of exploitation: once the optimal action is discovered, an agent can stop sampling alternative actions as it is not important to know their precise values. In contrast, in lean environments, choosing the optimal action typically lowers its value (due to sparse rewards), to the point that it can drop below those of even more suboptimal actions. Higher levels of exploration are therefore needed to accurately learn the value of the worse options in order to avoid them more reliably in the long run. Moreover, while in computer science applications one might be able to simply tune hyperparameters of an RL model for a given environment, ecologically, an agent cannot know whether it is in a rich or lean environment in advance.

In this paper, we investigate the utility of nonlinear basal ganglia opponency for adaptive behavior in rich and lean environments. We propose a new model, OpAL*, which (as observed empirically; ***Hamid et al.*** (***2015***)) dynamically adapts its dopaminergic state online as a function of learned reinforcement statistics of the environment. Specifically, OpAL* modulates its dopaminergic states in proportion to its estimates of "environmental richness", leading to dynamically evolving high DA

motivational states in rich environments and lower DA states in lean environments with sparse rewards. This dynamic modulation amplifies the D1 or D2 actor most well suited to discriminate amongst benefits or costs of choice options for the given environment, akin to an efficient coding strategy. We compared the performance of OpAL* to several baseline models (including alternative formulations of striatal opponency; ***Möller and Bogacz*** (***2019***)), to specifically test the need for the biological mechanisms in support of adaptive behavior. We find that OpAL* optimizes action selection across a range of environments with varying reward rates and complexity levels, and across a wide range of parameter settings. This advantage depends on opponency, nonlinearity, and adaptive DA modulation and is most prominent in lean environments, an ecologically probable environment which requires more adaptive navigation of explore-exploit as outlined above. OpAL* also addresses limitations of the original OpAL model highlighted by ***Möller and Bogacz*** (***2019***), while retaining key properties needed to capture a range of empirical data and afford the normative advantages. Finally, we apply OpAL* to capture a range of empirical data across species, including how risk preference changes as a function of D2 MSN activity and manipulations that are not explainable by monolithic RL systems even when made sensitive to risk (***Zalocusky et al., 2016***). In humans, we show that OpAL* can reproduce patterns in which dopaminergic drug administration selectively increases risky choices for gambles with potential gains (***Rutledge et al., 2015***). Moreover, we show that even in absence of biological manipulations, OpAL* also accounts for recently described economic choice patterns as a function of environmental richness. In particular, we simulate data showing that when offered the very same safe and risky choice option, humans are more likely to gamble when that offer had been presented in a the context of a richer reward distribution (***Frydman and Jin, 2021***). Taken together, our simulations provide a clue as to the normative function of the biology of RL which differs from that assumed by standard models and gives rise to variations in risky decision making.

OpAL overview

Before introducing OpAL*, we first provide an overview of the original OpAL model (***Collins and Frank, 2014***), an algorithmic model of the basal ganglia whose dynamics mimic the differential effects of dopamine in the D1/D2 pathways described above. OpAL is a modified "actor-critic" architecture (***Sutton and Barto, 2018***). In the standard actor-critic, the critic learns the expected value of an action from rewards and punishments and reinforces the actor to select those actions that maximize rewards. Specifically, after selecting an action (a), the agent experiences a reward prediction error (δ) signaling the difference between the reward received (R) and the critic's learned expected value of the action ($V_t(a)$) at time t:

$$\delta_t = R_t - V_t(a) \tag{1}$$

$$V_{t+1}(a) = V_t(a) + \alpha \times \delta_t, \tag{2}$$

where α is a learning rate. The actor then selects actions based on their relative action propensities, using a softmax decision rule:

$$p(a) = \frac{e^{Act_t(a)}}{\sum_{i \in A} e^{Act_t(i)}}, \tag{3}$$

where Act values are updated as a function of reward prediction errors in the critic, such that the agent selects those actions that yield the most frequent positive RPEs. OpAL is distinguished from a standard actor-critic in two critical ways, motivated by the biology summarized above. First, it has two separate opponent actors: one promoting selection ("Go") of an action a in proportion to its relative benefit over alternatives, and the other suppressing selection of that action ("NoGo")

in proportion to its relative cost (or disappointment). [1] Second, the update rule in each of these actors contains a *three-factor Hebbian rule* such that weight updating is proportional not only to learning rates and RPEs (as in standard RL) but is also scaled by G_t and N_t themselves. In particular, positive RPEs conveyed by phasic DA bursts strengthen the G (D1) actor and weaken the N (D2) actor, whereas negative RPEs weaken the D1 actor and strengthen the D2 actor.

$$G_{t+1}(a) = G_t(a) + \alpha_G G_t(a) \times \delta_t \tag{4}$$

$$N_{t+1}(a) = N_t(a) + \alpha_N N_t(a) \times -\delta_t \tag{5}$$

where α_G and α_N are learning rates controlling the degree to which D1 and D2 neurons adjust their synaptic weights with each RPE. We will refer to these G_t and N_t terms that multiply the RPE in the update as the "Hebbian term", because weight changes grow with activity in the corresponding G and N units. As such, the G weights grow to represent the benefits of candidate actions (those that yield positive RPEs more often, thereby making them yet more eligible for learning), whereas the N weights grow to represent the costs or likelihood of disappointment (those that yield negative RPEs more often).

The resulting nonlinear dynamics capture biological plasticity rules in neural networks, where learning depends on dopamine (δ_t), presynaptic activation in the cortex (the proposed action a is selectively updated), and postsynaptic activation in the striatum (G_t or N_t) (***Frank, 2005***; ***Wiecki et al., 2009***; ***Beeler et al., 2012***; ***Gurney et al., 2015***; ***Frémaux and Gerstner, 2016***; ***Reynolds and Wickens, 2002***). Incorporation of this Hebbian term prevents redundancy in the D1 vs D2 actors and confers additional flexibility, as described in the next section. It is also necessary for capturing a variety of behavioral data, including those associated with pathological aberrant learning in DA-elevated and depleted states, whereby heightened striatal activity in either pathway amplifies learning that escalates over experience (***Wiecki et al., 2009***; ***Beeler et al., 2012***; ***Collins and Frank, 2014***).

For action selection (decision-making), OpAL combines together $G_t(a)$ and $N_t(a)$ into a single action value, $Act(a)$, but where the contributions of each opponent actor are weighted by corresponding gains β_g and β_n.

$$Act_t(a) = \beta_g G_t(a) - \beta_n N_t(a) \tag{6}$$

$$\beta_g = \beta(1 + \rho) \tag{7}$$

$$\beta_n = \beta(1 - \rho) \tag{8}$$

$$\tag{9}$$

Here, ρ reflects the (*dopaminergic state*) controlling the relative weighting of β_g and β_n, and β is the overall softmax temperature. Higher β values correspond to higher exploration, while $\beta = 0$ would generate random choice independent of learned values. When $\rho = 0$, the dopaminergic state is "balanced" and the two actors G and N (and hence, learned benefits and costs) are equally weighted during choice. If $\rho > 0$, benefits are weighted more than costs, and vice-versa if $\rho < 0$. While the original OpAL model assumed a fixed, static ρ per simulated agent to capture individual differences or pharmacological manipulations, below we augmented it to include the contributions of dynamic changes in dopaminergic state, so that ρ can evolve over the course of learning to optimize choice.

Nonlinear OpAL dynamics support amplification of action-value differences

After learning, G and N weights correlate positively and negatively with expected reward, with appropriate rankings of each action preserved in the combined action value Act (***Collins and Frank,***

[1] For clarity, "benefits" and "costs" are evaluations relative to the critic's expectation. The exact numeric value is not interpretable. Rather, high benefits (G) convey that an action is better than expected more often; high costs (N) convey that an action more often disappoints relative to the critic's expectations.

2014). Nevertheless, the Hebbian term induces nonlinear dynamics in the two actors such that they are not redundant and instead specialize in discriminating between different reward probability ranges (Figure 1). While the G actor shows greater discrimination among frequently rewarded actions, the N actor learns greater sensitivity among actions with sparse reward. Note that if G and N actors are weighted equally in the choice function ($\rho = 0$), the resultant choice preference is invariant to translations across levels of reward, exhibiting identical discrimination between a 90% and 80% option as it would between a 80% and 70% option. This "balanced" OpAL model therefore effectively reduces to a standard non-opponent RL model, but as such, fails to capitalize on the underlying specialization of the actors (G and N) in ongoing learning. We considered the possibility that such specialization could be leveraged dynamically to amplify a given actor's contribution when it is most sensitive, akin to an "efficient coding" strategy (***Frydman and Jin, 2021***).

OpAL*

Given the differential specialization of G vs N actors, we considered whether the critic's online estimation of environmental richness (reward rate) could be used to control dopaminergic states (as seen empirically; (***Hamid et al., 2015***; ***Mohebi et al., 2019***)). Due to its opponent effects on D1 vs D2 populations, such a mechanism would differentially and adaptively weight G vs N actor contributions to the choice policy. To formalize this hypothesis, we constructed OpAL*, which uses an online estimation of environment richness to dynamically amplify the contribution of the actor theoretically best specialized for the environment type.

To provide a robust estimate of reward probability in a given environment, OpAL* first replaces the standard critic with a Bayesian critic (so that value estimates are updated in proportion to uncertainty; see ***Franklin and Frank*** (***2015***) for possible striatal implementations of Bayesian learning via cholinergic interactions with dopamine in spiny cells). As such, the probability of reward for a given action $\hat{p}(r, a)$ is represented by a beta distribution rather than a point estimate. The critic then generates a prediction error as the obtained reward relative to the expected value of an action using the mean of the beta distribution, $\hat{p}_t(r, a)$, multiplied by the magnitude of reward R_{mag} and loss L_{mag}. Unless otherwise noted, simulations in this paper use $R_{mag} = 1$ and $L_{mag} = 0$.

$$\alpha^c_{t+1}(a) = \alpha^c_t(a) + R_t \tag{10}$$

$$\beta^c_{t+1}(a) = \beta^c_t(a) + (1 - R_t) \tag{11}$$

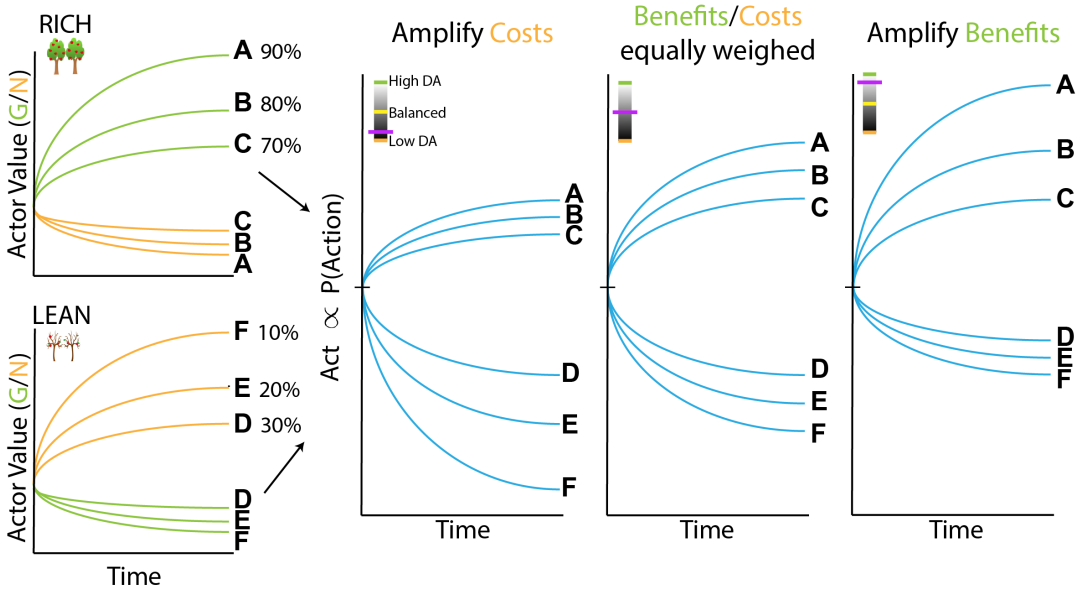
$$X \sim \text{Beta}(\alpha^c_t(a), \beta^c_t(a)) \tag{12}$$

$$\hat{p}_t(r, a) = E[X] \tag{13}$$

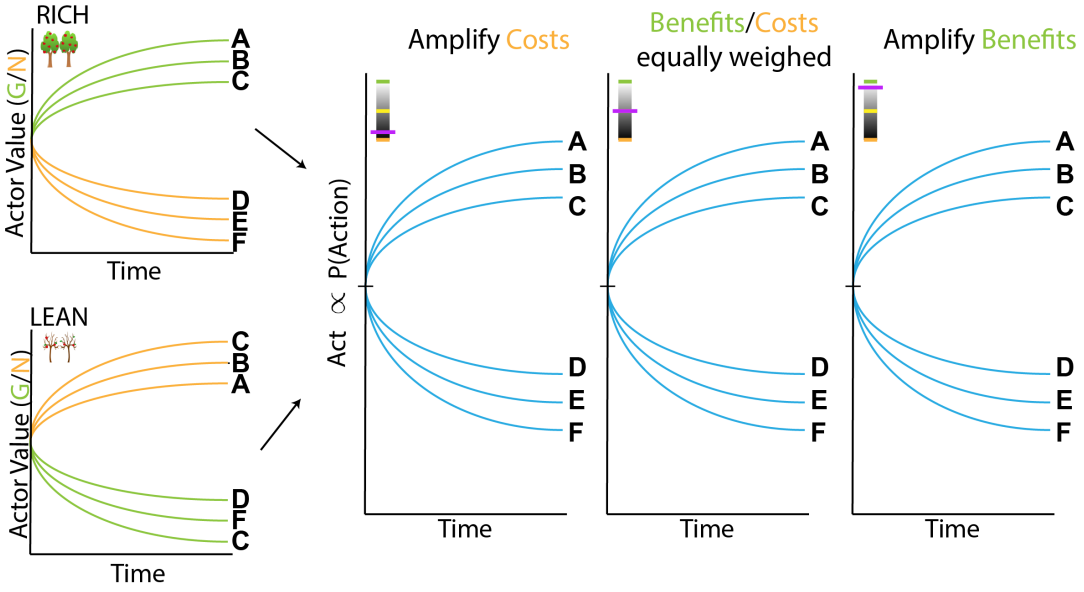
$$V_t(a) = R_{mag} \times \hat{p}_t(r, a) + L_{mag} \times (1 - \hat{p}_t(r, a)), \tag{14}$$

where α^c and β^c are hyperparameters of the beta distribution. This prediction error is then used to train the G and N actors, as noted above. OpAL* also uses a beta distribution to estimate $\hat{p}_t(r)$ for the environment as a whole (i.e., over all actions), or "state-value", by combining the alpha and betas from each action. The dopaminergic state ρ is then increased when $\hat{p}_t(r) > .5$ (*rich environment*), and decreased when $\hat{p}_t(r) < .5$ (*lean environment*). To ensure that dopaminergic states accurately reflect environmental richness, we apply a conservative rule to modulate ρ only when the critic is sufficiently "confident" that the reward rates are above or below 0.5, that is, we take into account not only the mean but also the variance of the beta distribution, parameterized by ϕ (Equation 16). This process is akin to performing inference over the most likely environmental state to guide DA.[2] Lastly, a constant k controls the strength of the modulation (Equation 17)

[2] One can adjust DA without the conservative inference process but there is a cost to misestimation of environmental richness that can arise due to stochasticity in any given environment, which can lead to reliance on the wrong actor; see Appendix. Although we focus on the Bayesian implementation here, other heuristics for achieving the same desideratum can be applied, for example waiting a fixed number of trials before changing the dopaminergic state using a standard RL critic. However, using a beta distribution (whose mean implicitly incorporates uncertainty) and explicitly adapting according to the distributions' standard deviation isolates whether any differences in performance between OpAL* and a baseline model with fixed dopamin-



(a) Schematic of OpAL dynamics with three-factor Hebbian term. Nonlinear weight updates due to Hebbian factor leads to increasing discrimination between high reward probability options in the G actor and between low reward probability options in the N actor. For intermediate dopamine states (G and N actors are balanced), there is equally sensitive to differences in reward probability across the range of rich and lean environments. For high dopamine states ($\beta_g > \beta_n$), the action policy emphasizes differences in benefits (as represented in the D1/"G" weights), whereas in low dopamine states ($\beta_g < \beta_n$), the action policy emphasizes differences in costs (as represented in the D2/"N" weights).



(b) Schematic of OpAL dynamics without three-factor Hebbian term. Removing nonlinear term in OpAL confers redundancy in G/N weights, which are anticorrelated and thus cannot be leveraged to promote specialization in different ranges.

Figure 1. Changes in dopaminergic state (represented by the purple indicators) affect the policy of OpAL due to its nonlinear and opponent dynamics. OpAL* hypothesizes that modulating dopaminergic state by environmental richness is a normative mechanism for flexible weighting of these representations.

$$Y \sim \text{Beta}(\alpha_t^{\text{state}}, \beta_t^{\text{state}}) \tag{15}$$

$$S = \begin{cases} 1 & \text{if E[Y] - } \phi \text{ std(Y) > .5} \\ 1 & \text{if E[Y] + } \phi \text{ std(Y) < .5} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

$$\rho_t = S(E[Y] - .5)k, \qquad k \geq 0 \tag{17}$$

Choice To accommodate varying levels of k and to maintain biological plausibility, the contribution of each actor is lower-bounded by zero – that is, G and N actors can be suppressed but cannot be inverted (firing rates cannot go below zero), while still allowing graded amplification of the other subpopulation.

$$Act_t(a) = \beta_g G_t(a) - \beta_n N_t(a) \tag{18}$$

$$\beta_g = \beta \max(0, 1 + \rho_t) \tag{19}$$

$$\beta_n = \beta \max(0, 1 - \rho_t) \tag{20}$$

Normalization and annealing The original three-factor Hebbian rule presented in ***Collins and Frank*** (***2014***) approximates the learning dynamics in the neural circuit models needed to capture the associated data and also confers flexibility as described above. However, it is also susceptible to instabilities under particular circumstances, as highlighted by ***Möller and Bogacz*** (***2019***). Specifically, because weight updating scales with the G and N values themselves, one can engineer a series of outcomes that can cause the weights to decay rapidly toward 0 (see Appendix). To address this issue, OpAL* introduces two additional modifications based on both functional and biological considerations. Firstly, we apply a transformation to the actor prediction errors such that they are normalized by the range of available reward values (see ***Tobler et al.*** (***2005***) for evidence of such normalization in dopaminergic signals). Secondly, the actor learning rate is annealed across time (see ***Franklin and Frank*** (***2015***) for a plausible circuit mechanism allowing striatal learning to stabilize across time, while remaining flexible to change points). These modifications improve the robustness of OpAL* and ensure that the actor weights are well-behaved, while preserving the key Hebbian features of OpAL (which, as shown below, are needed for its normative advantages). For a full discussion on these modifications, see Appendix.

$$G_{t+1}(a) = G_t(a) + \alpha(t) G_t(a) f(\delta_t) \tag{21}$$

$$N_{t+1}(a) = N_t(a) + \alpha(t) N_t(a) f(-\delta_t) \tag{22}$$

$$\alpha(t) = \frac{\alpha}{1 + t/T} \tag{23}$$

$$f(x) = \frac{\delta_t}{R_{mag} - L_{mag}} \tag{24}$$

Results

Robust advantages of adaptively modulated dopamine states

We hypothesized that OpAL* confers adaptive flexibility especially when an agent does not have information about the statistics of a novel environment and thus the agent cannot choose its hyperparameters accordingly. In this section, we therefore characterize the robustness of OpAL* advantages across a large range of parameter settings. We then explore how such advantages scale

ergic states were a result of dopamine modulation rather than an ineffective use of the critic (e.g., waiting too few trials) or a suboptimal critic (e.g., poorly tuned learning rate).

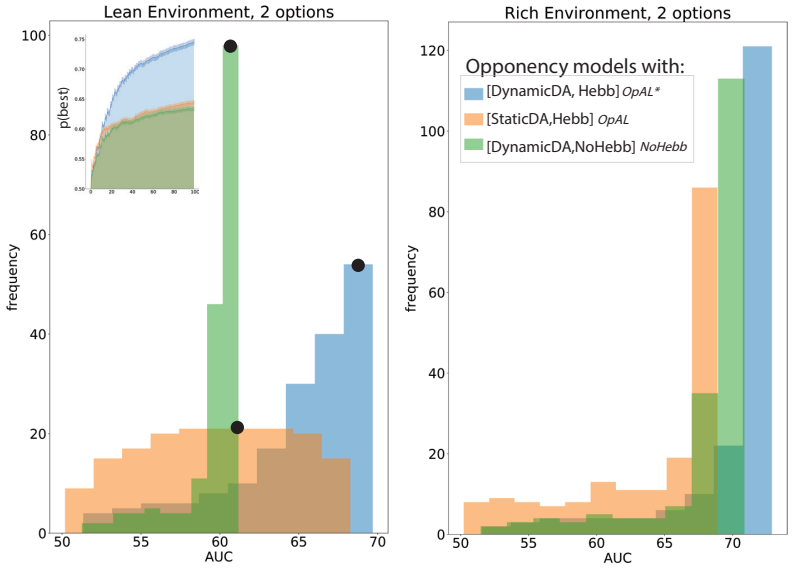


Figure 2. Biological mechanisms incorporated in OpAL* support robust advantages over control models across parameter settings in a reward rich environment (80% vs 70% two-armed bandit) and a lean, sparse reward environment (30% vs 20% two-armed bandit). Advantages over balanced OpAL model indicate need for dynamic dopamine modulation. Advantages over No Hebb model indicate the need for the nonlinear three-factor Hebbian rule (found in Equations 21 and 22). Together, advantages over both control models also indicate need for opponency, particularly given redundancy in G and N weights in the NoHebb model (see text, Figs 1 and 6, and additional comparisons to Q learner below). Figure shows area-under-the-curve (AUC) histograms of average learning curves for all parameters in a grid sweep. Black dots (left figure) indicate example AUC values which correspond to the shaded region under the respective learning curve (average softmax probability of selecting the best option, 30%) for each respective model. See Parameter grid search for more details.

with complexity in environments with increasing number of choice alternatives. In the subsequent section, we illustrate the mechanisms of such effects.

To specifically assess the benefit of adaptive dopaminergic state modulation, we first considered rich (80% vs. 70%) and lean (30% vs. 20%) 2-armed bandit environments. We compared OpAL* to two control models to establish the utility of the adaptive dopamine modulation (which was not a feature of the original OpAL model), and to test its dependence on nonlinear Hebbian updates. More specifically, the OpAL model equally weights benefits and costs throughout learning ("$\rho = 0$"); as such, any OpAL* improvement would indicate an advantage for dynamic dopaminergic modulation. [3] The No Hebb model reinstates the dynamic dopaminergic modulation but omits the Hebbian term in the three factor learning rule (Equations 21, 22). This model therefore serves as a test as to whether any OpAL* improvements depend on the underlying nonlinear actor weights produced by the three-factor Hebbian rule. The No Hebb model also serves to compare OpAL* to more standard single-actor RL models; removable of the Hebbian term renders each actor redundant, effectively a single-actor model (See Section Mechanism for more detail). Improvement of OpAL* relative to the No Hebb model would therefore suggest an advantage of OpAL* over standard actor-critic models (we also test OpAL* against a standard Q-learner below). Importantly, models were equated for computational complexity, with modulation hyperparameters (ϕ and k) of dynamic DA models (OpAL* and No Hebb) held constant (see Methods).

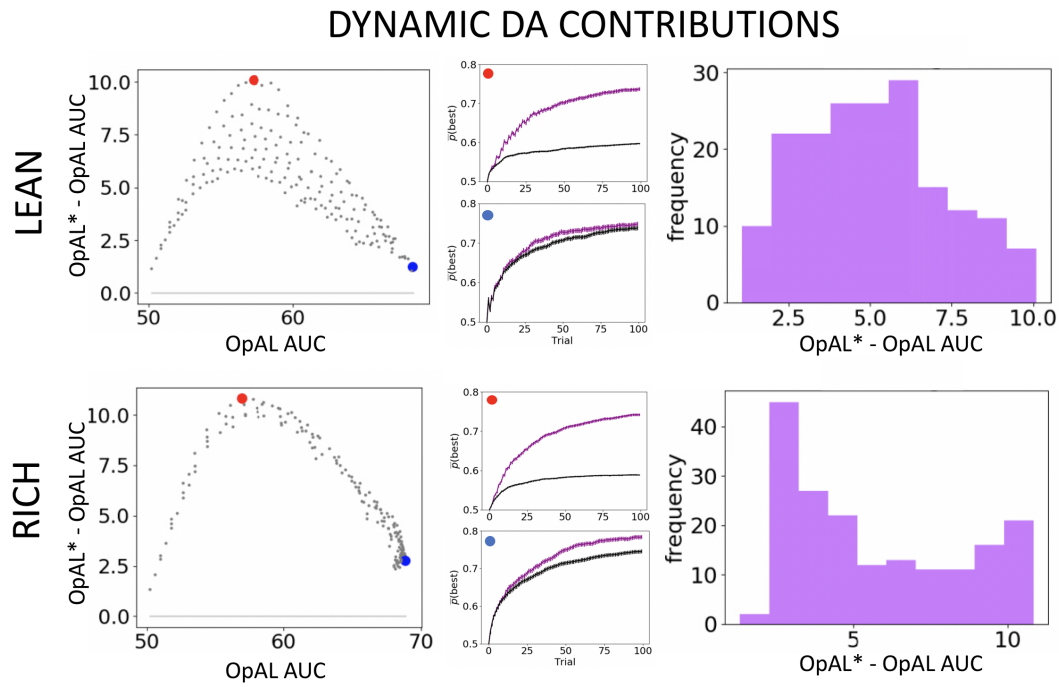
Following an initial comparison in the simplest two choice learning situation, we tested whether OpAL* advantages may be further amplified in more complex environments with multiple choice alternatives. We introduced additional complexity into the task by adding varying numbers of alternative suboptimal actions (e.g., an environment with four actions with probability of reward 80%, 70%, 70%, and 70%). Results were similar for average learning curves and average reward curves; we focus on average learning curves as they are a more refined, asymptotically sound measure of normative behavior.

We begin with the results of the two-choice paradigm (80%/70% or 30%/20%). For each parameter combination, we calculated the area under the curve (AUC) of the learning curves and then plotted histograms of these AUCs across all parameter sets (Figure 2). The first result that is apparent is that OpAL* outperforms its balanced OpAL control ($\rho = 0$) especially in the lean (sparse reward) environment. The mean of the OpAL* distribution is shifted higher and the shape is skewed rightward, due to selective improvement of moderate performing models (Figure 3a). The improvement is less dramatic in the rich environment, but is still evident and the distributions and more condensed around the peak, indicating robustness. Moreover, note that these improvements over balanced OpAL provide a lower bound estimate on the advantages of adaptive modulation, given that using any other fixed $\rho \neq 0$ would perform worse across environments: models with $\rho > 0$ perform very poorly in lean environments and those with $\rho < 0$ perform very poorly in rich environments (see Appendix). Finally, the non-Hebbian model performs dramatically worse in the lean environment in comparison to both OpAL* and the OpAL model, suggesting that OpAL* advantages require nonlinear Hebbian updates. Furthermore, we see here that OpAL* outperforms the best performing control within each environment alone.

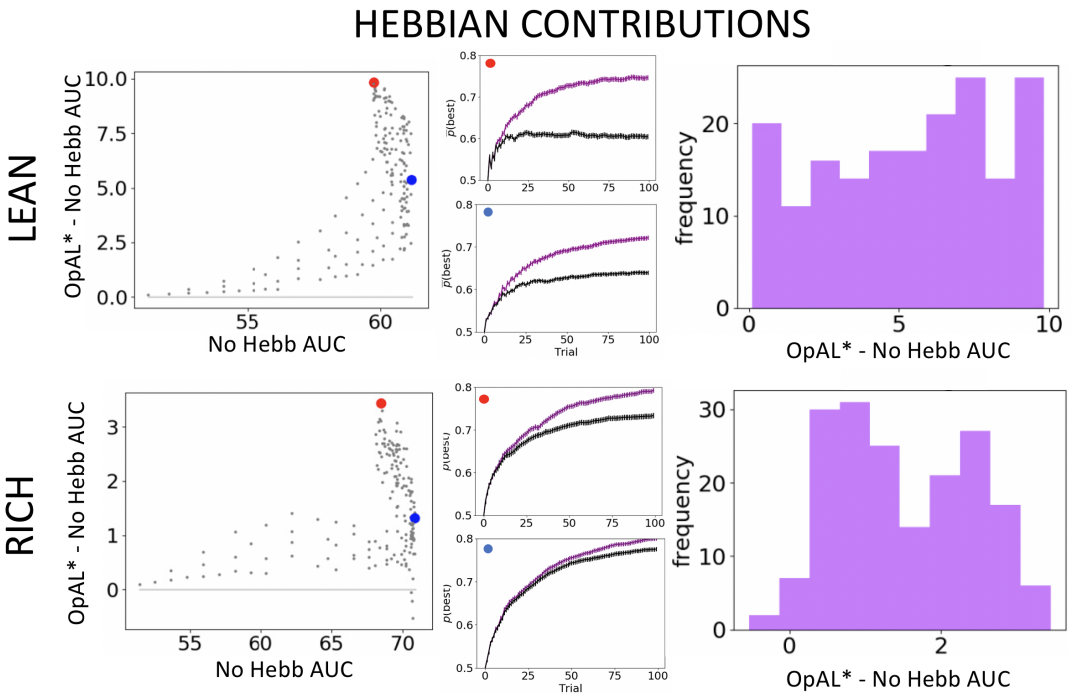
Overall, these results show an advantage for dynamic dopaminergic states as formulated in OpAL* when reward statistics of the environment are unknown. This advantage is particularly prominent in the lean (sparse reward) environment, which is computationally more challenging and ecologically more realistic than the rich environment. Crucially, dynamic dopaminergic state leverages the full potential of opponency ONLY when combined with three-factor Hebbian learning rules, as demonstrated by OpAL*'s advantage over the No Hebb model.

To statistically investigate where dopaminergic modulation was most advantageous, we per-

[3] In our simulations, the OpAL model includes the annealing and normalization additions as discussed in Section OpAL*. While these features were not present in the original version presented in ***Collins and Frank* (*2014*)**, we found they are necessary to address pathological behavior as discussed in Section OpAL* and in the Appendix. The crucial distinction we emphasize between OpAL and OpAL* is the non-dynamic versus dynamic adaptation of DA, respectively.

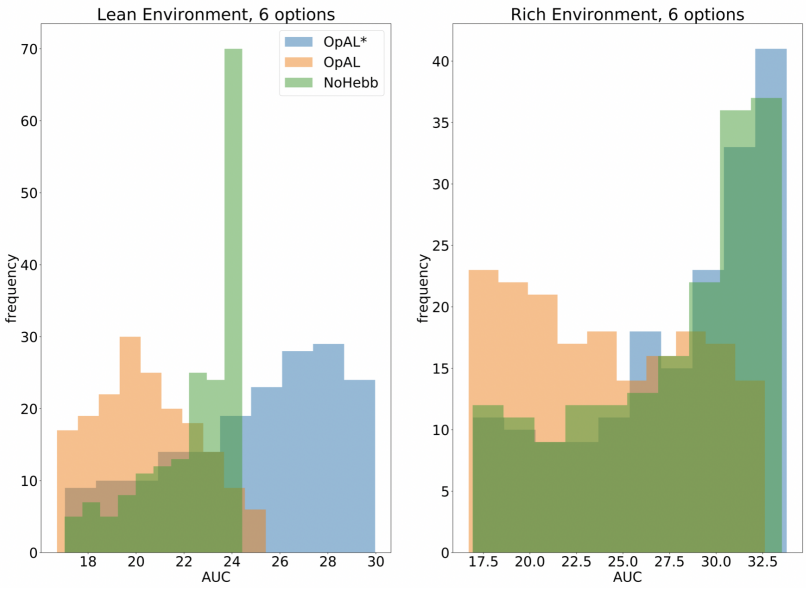


(a) OpAL* improves upon a control model which lacks dynamic modulation (OpAL, $\rho = 0$), with largest improvement for moderately performing parameters. Left, Each point represents a single parameter combination and its difference in learning curve AUCs in OpAL* compared to the OpAL model. Center, average learning curves of the parameter setting which demonstrates the best improvement of OpAL* over the OpAL model (indicated by the red dot) and the parameter setting with the best OpAL model performance (indicated by the blue dot). Right, Histogram of the difference in average learning curve AUCs of the two models with equated parameters. All results also correspond to Figure 2.

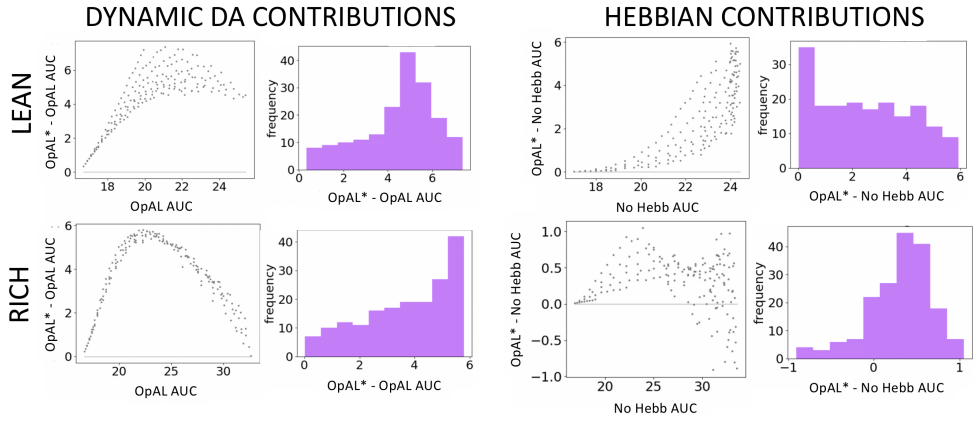


(b) Dynamic DA modulation is insufficient to induce performance advantage without three-factor Hebbian learning (No Hebb). Comparison descriptions analogous to the above. All results also correspond to Figure 2.

Figure 3. Parameter level comparison of OpAL* to a OpAL model and OpAL* to the No Hebb model across a range of plausible parameterss. Results of two-armed bandit environments – rich (80% vs. 70%) or lean (30% vs. 20%) – for 100 trials. See Parameter grid search for further details of methods.



(a) AUC Histogram of average learning curve for various parameter settings in high complexity.



(b) Parameter level comparison of OpAL* to control models in high complexity.

Figure 4. OpAL* robustly outperforms control models in high complexity environments, with lean environments showing the greatest advantage. Models completed a 6-armed bandit task (with only one optimal action) for 100 trials. See Parameter grid search for detailed analysis methods.

formed one sample t-tests where the null was zero on the difference between the AUC of OpAL* and each control model for every parameter combination over several time horizons (50, 100, 250, and 500 trial; see Appendix for details). OpAL* outperformed its OpAL ($\rho = 0$) control and the non-Hebbian version across all time horizons ($p's < 1.0e^{-}47$). We can visualize these statistics plotted according to the AUC of the control model as well as the frequency of the AUC differences (Figure 3). Interestingly, OpAL* advantages over the OpAL model show an inverted-U relationship, whereby improvements are most prominent for mid-performing parameter combinations. In contrast, improvements relative to the No Hebb model (Figure 3b) are most prominent for high performing baseline parameter combinations.

OpAL* advantages grow with environmental complexity

We next explored these effects in progressively more complex environments by increasing the number of available choice alternatives, across several time horizons (50, 100, 250, and 500 trials). Each complexity level introduced an additional suboptimal action to the rich or lean environment. For example, a complexity level of 4 for the lean environment consisted of four options: a higher

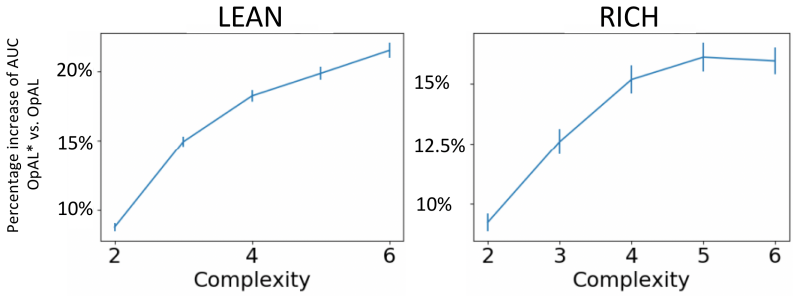


Figure 5. Advantage of dynamic dopaminergic modulation of OpAL* grows with complexity. Complexity corresponds to the number of bandits available in the environment (e.g. a 2-armed bandit, which data point corresponds to Figures 2 and 3, or a 6-armed bandit, which data point corresponds to Figure 4). Values reported are the average percentage increase of OpAL* learning curve AUC when compared to a OpAL model with equated parameters. That is, we computed the difference in AUC of OpAL* and OpAL model learning curves for a fixed parameter normalized by the AUC of the balanced OpAL model. We then averaged this percentage increase over all parameters in the grid search. Results are shown for 100 trials of learning.

rewarding option (30% probability of reward) and three equivalent lower rewarding options (20% probability of reward each).

OpAL* outperformed the OpAL model (differences in AUCs, $p's < 2.0e^{-}60$) across all time horizons and complexity levels. OpAL* also outperformed the non-Hebbian version ($p's < 1.0e^{-}4$), except for the highest complexity rich environments (5 or 6 options) after 500 trials ($p's > 0.1$; OpAL* advantages were still significant for lower trial counts).

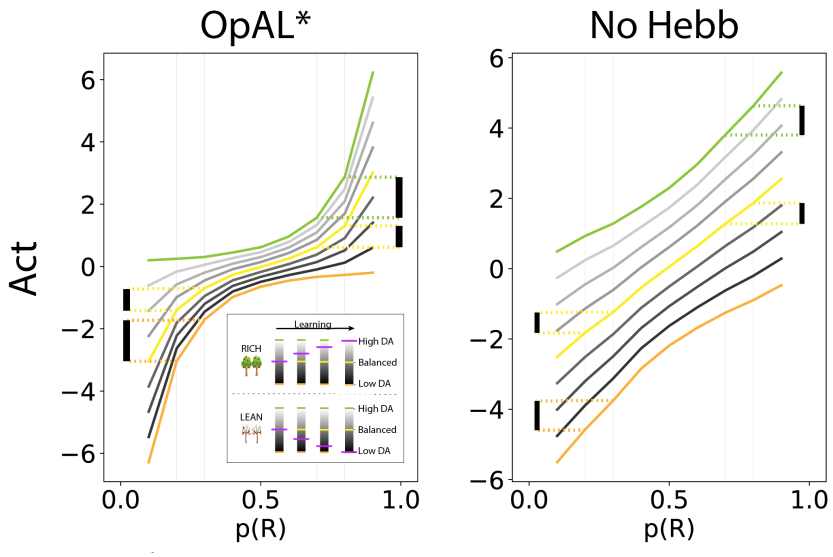
We can again visualize these results as AUC histograms for each model (Figure 4a) and as the AUC differences between matched parameters (Figure 4b). We visualize the highest complexity here for simplicity. As in the two-option results, the benefits of OpAL* are most evident in the lean environment (Figure 4a, left). OpAL* shows better performance across a range of parameters than control models. OpAL* is also the only model to achieve roughly equivalent performance in rich and lean environments in this parameter range. As noted in the introduction, standard RL models typically suffer in lean environments due to greater demands on exploration (see below for comparisons to more traditional RL models); these simulations show that OpAL* overcomes this robustness limitation and that its control models do not. OpAL* also shows less prominent, but nevertheless significant, advantages in the rich environment compared to the No-Hebb variant (Figure 4a, right), which can be visualized by the histogram of AUC differences between matched parameters (Figure 4b, bottom right). In lean environments, OpAL* improvements over the OpAL model were most evident for high performing parameter sets (positive trend in the scatter plot).

Finally, to assess the advantage of dynamic dopamine modulation, we quantified the OpAL* improvement over the balanced OpAL model as a function of complexity levels. Notably, OpAL* advantages grows monotonically with complexity, roughly doubling from low to high complexity levels (Figure 5).

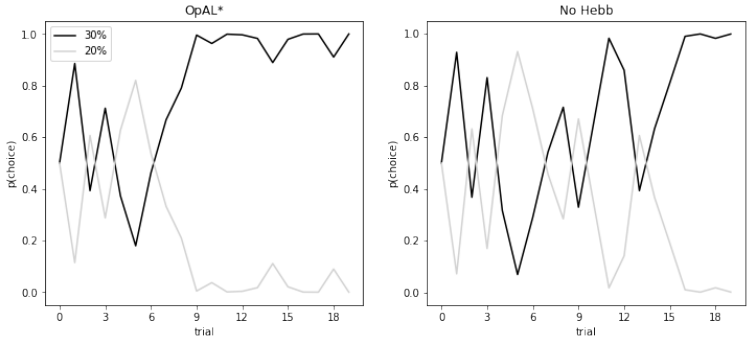
Mechanism

How does OpAL* confer such an advantage across environments? To illustrate the mechanism underlying this improvement, we considered two inter-related issues. The first issue concerns the dynamic leveraging of the nonlinearity in actor weights, and the second addresses the way in which the opponent mechanism navigates a particularly pernicious exploration/exploitation tradeoff that arises in lean environments.

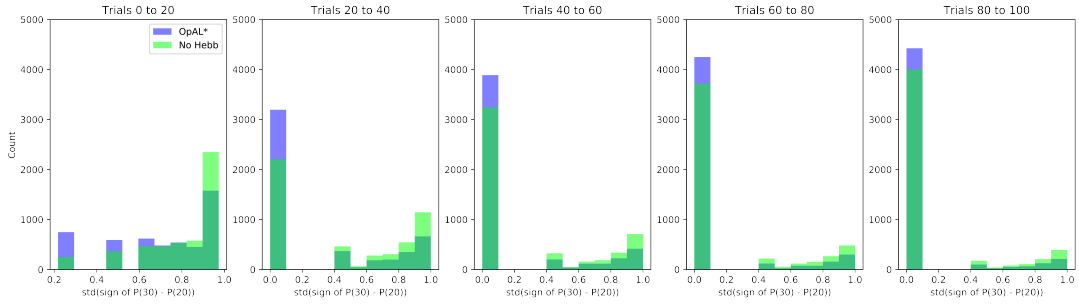
To observe the impact of nonlinearity, we plotted how Act values change as a function of reward probability and for different DA levels (represented as different colors, Figure 6a). While Act values increase monotonically with reward probability, the convexity in the underlying G and N weights (Fig 1a) gives rise to stronger Act discrimination between more rewarding options (e.g., 80% vs 70%)



(a) OpAL* capitalizes on convexity in actor weights in different environments. Left, Nonlinearity in OpAL* update rule induces convexity in Act values as a function of reward probability (due to stronger contributions of G weights with higher rewards, and stronger contributions of N weights with sparse reward). OpAL* dynamically adjusts its dopaminergic state over the course of learning as a function of its estimate of environmental richness (indicated by elongated, purple bars), allowing it to traverse different Act curves (high DA in green emphasises the G actor, low DA in orange emphasises the N actor). As such, OpAL* can differentially leverage convexity in G or N weights, outperforming a "balanced" OpAL model (in yellow) which equally weighs the two actors (due to static DA). Vertical bars show discrimination between 80% and 70% actions is enhanced with high dopamine state, whereas discrimination between 20% and 30% actions is amplified for low dopamine. Right, Due to redundancy in the No Hebb representations, policies are largely invariant to dopaminergic modulation during the course of learning. Act values are generated by presenting each model with a bandit using a fixed reward probability for 100 trials; curves are averaged over 5000 simulations.



(b) Linear weight updating (without Hebbian term) induces prolonged policy fluctuations in lean environments, which (after initial exploration) is overcome by OpAL*. See main text for explanation. Example curves from optimized parameters from Figure 7 using same random seeds across models.



(c) OpAL* demonstrates reduced policy fluctuations in early learning in sparse reward environments, as indexed by the sign of the difference between the probability of selecting the optimal action (30%) and the suboptimal action (20%). Higher standard deviation of this metric indicates more fluctuations in policy. Here we show the histogram of standard deviations of these signs across 5000 simulations (which correspond to Figure 7), illustrating higher exploration rates in the No Hebb model.

Figure 6. Overview of OpAL* mechanisms contributing to performance improvement relative to balanced OpAL and No Hebb variants.

with higher dopamine levels, and between less rewarding options (e.g., 30% vs 20%) with lower dopamine Levels. As the critic converges on an estimate of environmental richness, OpAL* can adapt its policy to dynamically emphasize the most discriminative actor (Figure 6a, left). In contrast, due the lack of nonlinearity, the No Hebb variant induces redundancy in the G and N weights and thus essentially reduces to a standard actor critic agent. As such, dopamine modulation does not change its discrimination performance across environments (Figure 6a, right).

If enhancing discrimination between action values improves performance, why could this not be achieved by simply increasing overall exploitation (e.g., softmax gain)? Note that the smooth Act curves discussed above depend on agents having already been exposed to reward probabilities (i.e., they were generated after learning). But as highlighted in the introduction, sparse reward environments typically require higher levels of exploration to accurately estimate action values. Indeed, in lean environments, repeated selection of the optimal action often leads to its value decreasing below that of suboptimal actions during early learning, causing the agent to switch to those suboptimal actions again until they become worse, and so on. This effect is evident in the No Hebb model, which is susceptible to substantial fluctuations in its policy in lean environments (Figures 6b and 6c). OpAL* overcomes this issue in two ways. First, opponency allows the non-dominant (here, G) actor to contribute early during learning (before N weights accumulate), thereby flattening initial discrimination and enhancing exploration. Second, the Hebbian nonlinearity ensures that negative experiences induce disproportional distortions in N weights for the most suboptimal actions after they have been explored (Figure 6a), thereby allowing the agent to more robustly avoid them (Figures 6b and 6c). By adapting its policy by environmental richness, OpAL* can dynamically leverage this specialization. In sum, OpAL* maintains specialized representations but can dynamically modulate when to use them to solve an explore-exploit tradeoff that is especially predominant in lean environments.

We conclude this section by considering whether the above discussion implies OpAL* might simply induce a more efficient change from exploration to exploitation across learning, as is sometimes considered in variants of standard RL. To diagnose whether dynamically modifying the softmax temperature alone is sufficient to improve robustness, we simulated a control variant in which both $_G$ and $_N$ were dynamically increased together, independent of the sign of ρ (Modulation model, see Appendix). Importantly, OpAL* outperformed the best-performing Modulation model across environments, and demonstrated notable improvement in lean environments. These simulations show that while dynamic changes in softmax temperature may be sufficient to improve performance in rich environments, the dynamic shift from one actor to another is integral to flexibility across both environments and especially for addressing the limitations of single actor models in lean environments. It is plausible that combining approaches (dynamic changes in both β modulation and ρ modulation) would show additional improvement. However, our focus is to investigate why dopaminergic modulation may be normatively useful and therefore such investigation is beyond the scope of this paper.

OpAL* outperforms alternative models with optimized parameters

The above simulations highlighted robustness of OpAL* advantages across large ranges of parameters using comparison models that are identical in every other respect. We next set out to compare OpAL* performance to other alternatives in the literature. For example, while the non-Hebbian model was the best control given every other aspect of it was identical to OpAL*, it still comprises an actor critic. Any claims that OpAL* confers an advantage should also be compared to the most common model-free RL agent, a Q-learner, which also maintains a single expected value across options. We also compared OpAL* to an alternative model of D1/D2 opponency ***Möller and Bogacz*** (***2019***). Given that we are now comparing models of different forms altogether, we optimized parameters in each case using gradient descent, thereby allowing each model to exhibit its best possible performance. Importantly, to equate degrees of freedom between OpAL and a standard Q-learner, DA modulation (ϕ, k) and annealing (T) parameters of OpAL* variants were held constant

during optimization procedures (see Methods).

To begin, we considered a standard Q learner by optimizing its learning rate and softmax temperatures, and optimized OpAL* over these same parameters, with both models tested across reward rich (80% vs. 70%) and lean (30% vs. 20%) 2-armed bandit environments (Figure 7). Because dopaminergic modulation may be most useful when the environment reward statistics are unknown, we optimized the parameters across *both* environments rather than optimized for each environment individually.

We first confirmed that OpAL* exhibits performance improvements over the Q learner in reward lean environments, and exhibits comparable performance in reward rich environments (see also ***Collins and Frank*** (***2014***)). Note that because only the learning rate and softmax temperature were optimized in OpAL*[4], these simulations provide a lower bound on the potential improvement for OpAL*.

To more specifically assess the benefit of adaptive dopaminergic modulation, we further compared OpAL* to three additional control models, where each model had its parameters optimized. The first model is an alternative to OpAL (but still opponent G/N model) proposed by ***Möller and Bogacz*** (***2019***). This model does not include the Hebbian term, but does include a different nonlinearity which (under some constraints) allows the G and N weights to converge to the mean expected payoffs and costs in the environment. This normative property serves as a useful comparison: once costs and benefits are known, an agent should be able to choose its policy to maximize reward. However, in actuality, the convergence to expected payoffs and costs in this model depends on having a constrained relationship between parameters optimized by *a priori* access to the distributions of rewards in the environment. Thus we hypothesized that OpAL* could more robustly optimize performance across environments with unknown statistics. Moreover, this control model serves as another test for the utility of the Hebbian term and the convexity of OpAL* G/N weights, as compared to the concave weights in ***Möller and Bogacz*** (***2019***). For completeness, we also include the other two OpAL control models from the previous section: the "balanced" ($\rho = 0$) OpAL model which lacks dynamic DA and the "No Hebb" OpAL model which omits the Hebbian term in the weight update. Because of the redundancy in G / N weights, the non-Hebbian model also serves as another baseline comparison for standard RL models, like Q learning, but with all other aspects of the model equated.

OpAL* outperformed all control models when each of them were optimized (Figure 7). Relative to the OpAL model, OpAL* adaptively modulated its choice policy to increase dopamine levels ($\rho > 0$) in rich environments, but to decrease dopamine levels ($\rho < 0$) in lean environments (See Figure 6a.) Indeed, performance advantages are especially apparent in reward lean environments, providing a computational advantage for low dopamine levels that can accentuate differences between sparsely rewarded options. Notably, performance advantages in lean environments depended on the Hebbian term. While other models (including "standard" RL) make qualitative predictions that performance should be significantly lower for lean than rich environments, OpAL* shows substantially improved performance in lean environments. In line with these qualitative patterns of OpAL*, rodents showed equally robust learning in rich environments (90% vs. 50 % bandit task) compared to lean environments (50% vs. 10% bandit task) ***Hamid et al.*** (***2015***) (see Figure 1d of that paper).

Finally, the model in ***Möller and Bogacz*** (***2019***) demonstrated poor across-environment performance, performing only slightly above chance in the rich environment. Results are not shown for this model as it was not intended to be optimized across diverse environments. (Indeed, as noted above, it can perform quite well in any given environment if its parameters are chosen carefully, ***Möller and Bogacz*** (***2019***)). We further compared a grid search for the ***Möller and Bogacz*** (***2019***) model and again found a detriment in performance relative to OpAL*, with sensitivity to small deviations from its optimal parameter settings in a particular environment. This sensitivity worsened

[4]Other parameters were only moderately hand-tuned for reasonable performance. Optimizing only the learning rate and softmax temperature for both models ensured that the searched parameter space for the models was well-matched.

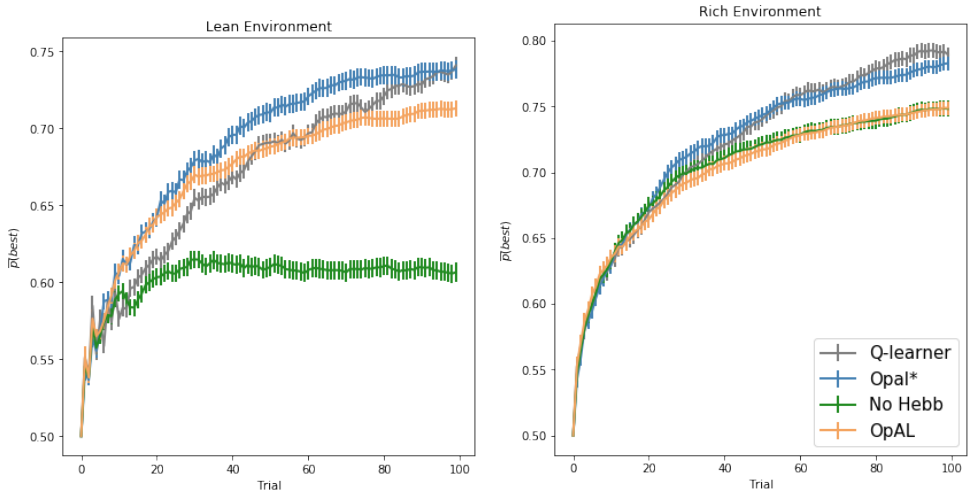


Figure 7. Comparison of OpAL* and various control models, each optimized for performance across rich and lean environments. Each curve is the mean softmax probability of selecting the best action over 5,000 simulations using the optimized parameters. Error bars are standard error of the mean. See Section Optimized Models in Materials and Methods for details on parameters and optimization procedure.

with complexity, and the model significantly underperformed relative to OpAL* across both environments. Thus, opponency itself is insufficient to capture the proposed advantages of OpAL*.

OpAL* adaptively modulates risk taking

Although the above analyses focused on learning effects, the adaptive advantages conferred by dopaminergic contribution were mediated by changes in the choice function (weighting of learned benefits vs costs), rather than learning parameters *per se*. We thus next sought to examine whether the same adaptive mechanism could also be leveraged for inferring when it is advantageous to make risky choices.

Models selected between a sure reward and a gamble of twice the value with unknown but stationary probability. The sure thing (ST) was considered the default reference point (***Kahneman and Tversky, 1979***), and gamble reward was encoded relative to that; that is, R_{mag} = +1 if gamble was won (gamble received an additional point relative to taking ST) or L_{mag} = -1 (loss of the ST). In high probability gamble states, the probability of reward was drawn uniformly above 50%; in low probability gamble states, probability of reward was drawn uniformly below 50%. Models were presented with the same gamble for 40 trials. The critic tracked the $\hat{p}(r)$ of the gamble and modulated ρ by its estimated expected value, as in Equations 15 through 17. G/N actors then tracked the action value of selecting the gamble. The probability of accepting the gamble was selected using the softmax choice function, such that accepting the gamble is more likely as the benefits (G) exceed the costs (N). Act definition can be found in Equation 18.

$$p(gamble) = \frac{1}{1 + e^{-Act(a)}}$$

As expected, OpAL* dynamically updated its probability of gambling and improved performance in comparison to the balanced OpAL, non-modulated model (Figure 8). In states with high probability ($> 50\%$), value modulation helped the model infer that the gamble was advantageous. In low probability gambles ($< 50\%$), value modulation aided in avoiding the gamble, which was unfavorable in the limit. Similar results were also obtained using a simpler (non-Bayesian) critic which learned only through a TD update rule.

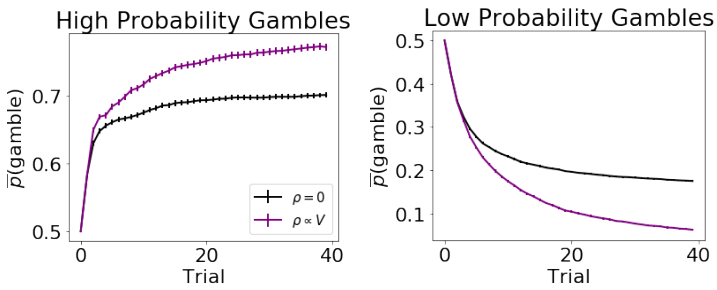


Figure 8. Dynamic dopamine modulation by estimated reward probability helps OpAL* decide when it is beneficial to gamble or to accept a sure reward. $\alpha_a = 1.0, \beta = 1.5$, annealing parameter $T = 10$, and modulation parameters $k = 20$ and $\phi = 0$. Results were averaged over 5,000 simulated states. Error bars are standard error of the mean. To limit variance, paired OpAL* and OpAL models were again given the same initial random seed.

OpAL* captures alterations in risky choice patterns across species

While all analyses thus far focused on normative advantages, the OpAL* model was motivated by biological data regarding the role of dopamine in modulating striatal contributions to cost/benefit decision making. We thus sought to examine whether empirical effects of DA and environmental richness on risky choice could be captured by OpAL* and thereby viewed as a byproduct of an adaptive mechanism. We focused on qualitative phenomena in empirical data sets that are diagnostic of OpAL* properties (and which should not be overly specific to parameter settings) and that could not be explained individually or holistically by other models. In particular, we consider impacts of optogenetic and drug manipulations of dopamine and stiratal circuitry in rodents and humans. We further show that OpAL* can capture economic choice patterns involving manipulation of environmental reward statistics rather than DA.

Striatal D2 MSN activity and reward history alters risky choice in rodents

Perhaps the most germane empirical study to OpAL since the original model was developed is that of (***Zalocusky et al., 2016***), who studied rodent risky choice as it is altered by reward history, dopamine manipulation, and striatal activity. Rats repeatedly chose between a certain option with a small reward or a gamble for larger reward whose expected value matched that of the certain option. Following unsuccessful gambles, they observed increased activity in D2-expressing medium spiny neurons (MSNs) in ventral striatum during subsequent decision periods. Recall that in OpAL*, reward history alters DA levels which in turn modulate activity in striatal MSNs and accordingly cost/benefit choice. In this case, a reduced recent reward history should reduce striatal DA, elevate D2 MSN activity, and thus promote choices that avoid costs. Indeed, Zalocusky et al. observed that animals were more likely to make a "safe" choice when D2 MSNs were stimulated during the choice period, and that endogenously, such safe choices were related to increased D2 activity and enhanced following unfavorable outcomes. Together, these results suggests an trial-to-trial adaptation of choice (rather than learning) driven by changes in D2 activity, akin to OpAL* mechanisms. Furthermore, such optogenetic stimulation effects were only seen in animals with a baseline preference for risk-seeking; risk-averse animals exhibited no change in behavior with the phasic manipulation.

Note first that these patterns of results are inconsistent with classical models in which striatal D2 activity is related only to motor suppression; here the impact of D2 activity is not to suppress actions altogether but instead to bias choices toward safe options. Instead, these results are consistent with OpAL* in which D2 activity is related to promoting actions with lowest perceived cost. Indeed, we found that this pattern of results align with the predictions of OpAL* but not alternative risk-sensitive models (see below).

As in previous sections, we encode gamble outcomes relative to the certain option: R_{mag} = +1

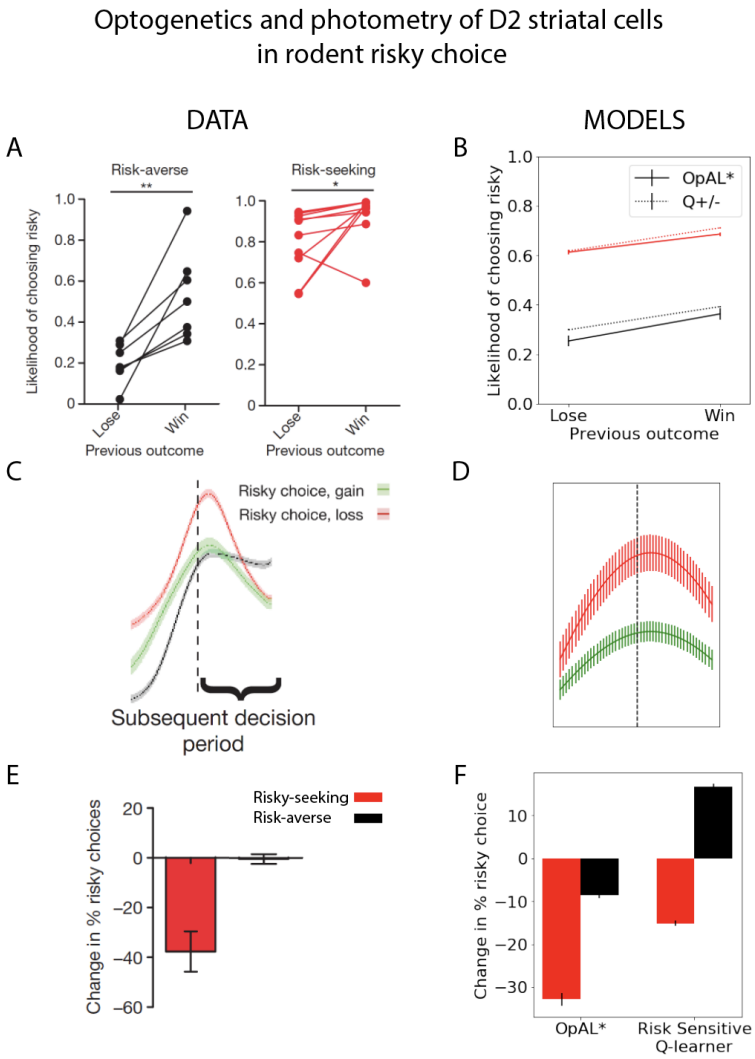


Figure 9. *Striatal D2 MSN activity and reward history alters risky choice in rodents*. Rodents repeatedly selected between a certain option with low magnitude of reward and a gamble with higher payout when successful.Left column: Modified figures from ***Zalocusky et al.*** (***2016***). Right column: Model simulations with OpAL* and Risk-sensitive RL (RSRL). **A,B.** Both risk-averse and risk-seeking animals are more likely to avoid a gamble after a gamble "loss" (failure to obtain the large reward). Both OpAL* and RSRL, a standard Q-learner with different learning rates for positive and negative prediction errors, can capture this trend, via changes in either choice function (D1 vs D2 MSN contributions) or learning rates, respectively. **C,D.** D2 MSN activity, measured via photometry, is larger after a gamble loss (red) than a gamble win (green) during the subsequent decision period. This pattern is reproduced in OpAL*, whereby D2 MSN activity is influenced by the product of the N weights and the adaptive β_n, which amplifies D2 MSN activity when dopamine levels are low. The simulation peak represents the average of this product after a loss or after a win, which is carried over to subsequent choices; error bars reflect SEM across simulations and dynamics before and after peak were generated by convolving the signal with a sinusoidal kernel for illustrative purposes. **E,F.** Optogenetic stimulation of D2 MSNs during the choice period induces risk-aversion selectively in risk-seeking rats. OpAL* captures this preferential effect by magnifying the effective D2 MSN activity and inducing avoidance primarily in risk-seeking agents. In contrast, RSRL predicts opposite patterns in risk-seeking and risk-averse animals. Parameters OpAL*: $\beta = 1.5, \alpha = 1., T = 20, k = 1.1, phi = 1.0$. Baseline ρ risk-seeking (0.85) and risk-averse (-0.75). Parameters RSRL: Risk-seeking $\alpha_+ = 0.3, \alpha_- = 0.1$; Risk-averse $\alpha_+ = 0.1, \alpha_- = 0.3, \beta = 1.5$). Since optogenetic effects were evident primarily during the choice period, we modeled this by changing the choice function in both models: in OpAL, trial-wise ρ values were decreased by 1.0 to mimic increased D2 MSN activity / decreased DA. In RSRL the choice function was altered by reducing β (to 0.01), leading to opposite directional effects in risk-seeking and risk-averse agents. Agents selected between a certain option and a 50/50 gamble with twice the payout for 100 trials.

if gamble was won or L_{mag} = -1. For OpAL*, the critic and actors operated as in section OpAL* adaptively modulates risk taking. G/N actors then tracked the value of selecting the gamble using the prediction error generated by the critic. As before, the probability of accepting the gamble was selected using the softmax choice function.

To simulate risk-seeking and risk-averse rats, we modified the baseline DA levels (ρ), holding all other parameters constant. Risk-seeking rats were modeled by higher levels of baseline ρ relative to those of simulations for risk-averse rats. To model phasic optogenetic stimulation, ρ values were decreased by a constant amount from this baseline.

We contrasted OpAL* to alternative models in which risky choice could be adapted. A popular model of dynamics in risky choice is called "risk-sensitive RL", in which an agent learns at different rates from positive and negative prediction errors:

$$Q(t+1) = Q(t) + \alpha_+ * PE, \text{ if } PE >= 0,$$
$$Q(t+1) = Q(t) + \alpha_+ * PE, \text{ if } PE < 0$$

,

where actions are selected using softmax function over Q values. If $\alpha_+ < \alpha_-$, an agent is more sensitive to risks in its environment. This formulation has been useful for characterizing asymmetric impacts of dopamine bursts and dips ***Frank et al.*** (***2007a***); ***Niv et al.*** (***2012a***), but focuses on learning rather than changes in choice functions. Because the effective manipulations on risky choice were made during the choice period rather than outcome, learning rate manipulations alone could not capture the effects. However, it is possible that DA or D2 manipulations can affect choice in simple RL models via simple changes to the overall softmax temperature, as assumed by many models (***FitzGerald et al., 2015***; ***Cinotti et al., 2019***; ***Eisenegger et al., 2014***; ***Lee et al., 2015***; ***Humphries et al., 2012***). We thus allowed the RSRL model to exhibit changes in risky choice by manipulating softmax gain accordingly, whereby D2 stimulation would mimic low DA levels and hence lower gain.

We found that both OpAL* and RSRL accounted for the decrease in gamble choices after gamble losses relative to wins, but generated opposing predictions for decision-period manipulation of D2-expressing neurons. While OpAL* predicts a decrease in riskiness in both risk-seeking and risk-averse rats (but more strongly in risk-seeking rats), RSRL predicts a decrease in riskiness in risk-seeking rats but an *increase* in riskiness in risk-averse rats. The reason for this effect is simply that a change in softmax gain leads to reduced exploitation, and thus drives both groups toward random selection. Thus the pattern of choice data is aligned with OpAL* but not with RSRL, or with classical models in which D2 activity inhibits choice altogether. These opposing predictions result from the *architecture* of OpAL* inspired by the biology– including opponency, Hebbian learning, and dynamic DA – rather than specific parameter values. Furthermore, OpAL* also captures the predicted relative activation of D2-expressing cells during the choice period following losses, due to changing DA levels ($\beta_n(t)$) and the learned cost of the gamble ($N(t)$), in line with Zalocusky's photometry data.

DA drug effects on risky decision-making and individual differences therein

We next focus on a human risky decision making paradigm manipulating DA levels (***Rutledge et al., 2015***). Participants were presented with interleaving trials of gain gambles (certain gain vs. potential greater gain or 0), loss gambles (certain loss vs. potential greater loss or 0), and mixed gambles (certain no reward vs. potential gain or potential loss). All gambles were successful with 50% probability. The study tested the effects of levodopa (L-DOPA), a drug which boosts dopamine release, on risky decision-making. The main impact of L-DOPA was to selectively amplify gambling on gain (but not loss or mixed) trials (Figure 10 A, left). This study also found that individual differences in this impact of drug on gambling correlated with effective drug dosage (Figure 10 B, left).

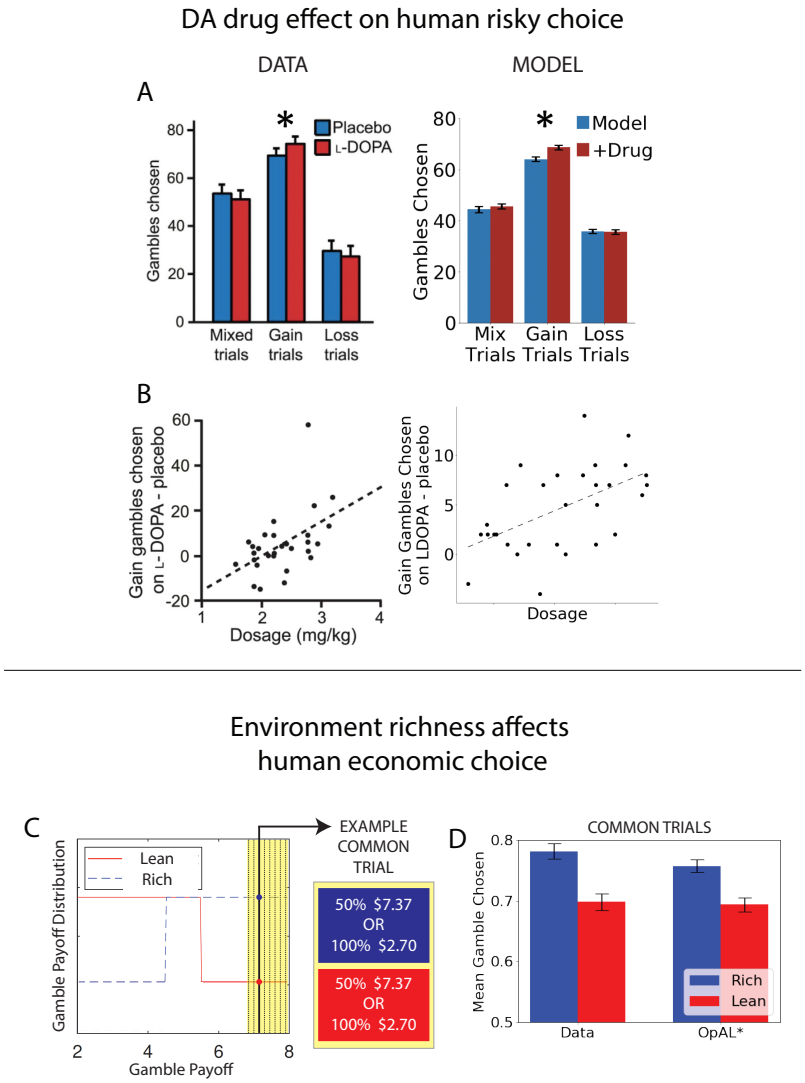


Figure 10. A,B. *DA drug effects on risky decision-making and individual differences therein.* OpAL* captures behavioral risk patterns of healthy participants on and off L-DOPA, a drug which boosts presynaptic DA. **A.** L-DOPA administration selectively increased risky choice in gain trials, where choice was between a sure reward and a 50% gamble for a larger reward, as compared to loss trials (sure loss vs gamble to avoid loss) or mix trials (in which gambles could result in gains or losses). **B.** These effects were larger for subjects with higher effective drug doses, Spearman's $\rho = 0.47, p < 0.01$. Left: Modified figures from ***Rutledge et al.*** (***2015***). Right: OpAL* simulations reproduce these selective effects. Spearman's $\rho = .50, p < .01$ To model individual differences in effective drug levels, for each pair of model on and off drug, d was drawn from a normal distribution centered at $.5$ with variance. $.25$. Parameters: $\beta = 1.5, k = 1.$. **C-D.** *Risky decisions are sensitive to environmental richness* In contrast to other empirical results discussed where dopamine pathways were directly manipulated, ***Frydman and Jin*** (***2021***) manipulated reward statistics of the payoffs in the environment, as in our normative simulations. Participants chose between a certain reward and a 50% gamble over two blocks. The distribution of payoffs in each block was either Rich (higher frequency of large magnitudes) or Lean (higher frequency of small magnitudes). Crucially, each block contained predetermined "common trials" where the payoff of both the gamble and certain option were fixed (e.g., an offer 50% $7.13 vs. 100% $2.70 was presented in *both* the Rich and Lean block). The key finding was that participants were more likely to gamble on these common trials when presented in the Rich context. OpAL* reproduces this pattern, due to adaptive ρ increasing DA levels and risk-tasking in the Rich block. Parameters:
$\alpha = 1., T = 10, \beta = 0.5, \phi = 1.0, k = 0.09$

The authors reported that the risk-seeking behavior with DA drugs was best described in terms of changes in a Pavlovian approach parameter. Here, we wished to see if the mechanisms introduced above within OpAL* with endogenous changes in dopaminergic state could replicate the pattern of results, thereby providing a normative interpretation.

We simulated 300 trials (100 gain gambles, 100 loss gambles, and 100 mixed gambles, randomly interleaved, as described in ***Rutledge et al.*** (***2015***)). Probability of gambling was determined as described above in the normative risky choice section, with gambles accepted as the benefits outweigh the costs relative to the ST. G and N actor values were explicitly set on each trial according to the instructed gamble and encoded relative to the certain option as in Section OpAL* adaptively modulates risk taking. This reduced the free parameters of OpAL* (no annealing or actor learning rate needed) while retaining its core features of DA reweighting the contributions of opponent representations during choice according to context.

While values and probabilities were explicitly instructed in the experiment, subjects nevertheless experienced the outcomes of each gamble. The OpAL* model assumes that they thus track the average value of offers across trials, such that a gain trial would elicit a positive dopamine deflection, given that its expected value is larger than that for mixed and loss trials. (As the authors note in discussing their findings, "in this task design, even the worst gain trial is better than the average trial and so likely inspires dopamine release.") We thus modeled the relative DA-state ρ proportional to the expected value of the current gamble offer, approximating how "rich" or "lean" the current offer was relative to all offers in the game. [5] (We formulate ρ proportional to value here, to be consistent with simulations in the above sections, but very similar results were obtained in a separate set of simulations in which ρ was modulated by RPE).

$$\rho(t) = .5 \times (\text{certain outcome}) + .5 \times EV(\text{gamble}) \tag{25}$$

To model L-DOPA, we hypothesized that it would boost positive RPEs via enhancement of evoked (phasic) DA release, as observed in vivo across species (***Voon et al., 2010***; ***Pessiglione et al., 2006a***; ***Qi et al., 2016***; ***Harun et al., 2016***). We assumed that L-DOPA amplified endogenous phasic release, which occurs when offers are better than usual (positive RPE). The effect dosage level was represented by d when the gamble had a positive value, as shown below.

$$\rho'(t) = \rho(t)(1 + d) \tag{26}$$

$$d \geq 0 \tag{27}$$

As hypothesized, OpAL* captured the selective effects of L-DOPA on gambling in gain trials. It also captured the overall proportion of gambles chosen for different trial types (Figure 10 A), as well as the correlation between effective dosage and difference in gambling on and off drug (Figure 10 B).[6] Furthermore, the Pavlovian model presented in ***Rutledge et al.*** (***2015***) would predict that gambling would occur for positive-RPEs even if the potential benefit of the gamble was not as high

[5] As gamble offers were explicit, removing uncertainty in trial richness, we omitted the parameter phi which modulated DA levels by degree of certainty in environmental richness, further reducing model complexity. OpAL*'s ability to capture shifting patterns of risky choice should thus be viewed as a byproduct of interacting opponent, nonlinear, and dynamic DA mechanisms rather than a result of high degrees of freedom.

[6] For clarification, ***Rutledge et al.*** (***2015***) highlighted that the drug effects appear "value-independent", whereas here we explicitly are changing risk sensitivity according to the interaction between drug and offer value. It is important to note, however, that their definition of value differs than that used to modulate dopaminergic state in these simulations. In ***Rutledge et al.*** (***2015***), value is defined as the advantage of the gamble, i.e., the difference between the expected value of the gamble and the sure reward. Here, we considered value to be the combined overall value of the offer presented, such that positive RPEs exist when values are greater than expected, and are in turn was modified by drug dosage. It is this component that captures the selective increase in gambling in gain trials. Note that the model does predict that such gambles would be yet more likely when the potential benefit of gambling is larger (i.e., when gains are particularly large) – but that this effect would also be present off drug. It is also possible that the value-independence in ***Rutledge et al.*** (***2015***) resulted from a ceiling effect for gambling in higher gain trials.

as the sure thing; OpAL* would only predict increased gambling if the benefits are greater than the sure thing.

Here, we have extended OpAL to account for risky decision-making by dynamically changing dopamine levels at choice proportional to the value of the current state/gamble offer. This accounted for findings of increase attractiveness of high-value risky options with the administration of L-DOPA (Figures 10 A). The model also accounted for individual differences of risk due to effective L-DOPA dosages (Figures 10 B). As highlighted in the previous section, these effects can normatively be explained as behavioral changes reflecting changes of inferred richness of current state. These results also suggest that individual differences in risk preference and sensitivity may be due to learned statistics of the world, casting these individual differences as deriving from an adaptive mechanism to an animal's or human's experience niche.

Risky decisions are sensitive to environmental richness: concordance with efficient coding models of economic choice

Thus far we have focused on data that are informative about the biological mechanisms (striatal opponency and DA modulation thereof) by which OpAL* supports adaptive behavior. But OpAL* also makes straightforward economic choice predictions that do not require biological manipulations. In particular, one way of conceptualizing OpAL* is that it serves as an efficient coding mechanism, by amplifying the actor that maximally discriminates between reward values in the current environment. If choice patterns concord with this scheme, one should be able to manipulate the environment and influence choice patterns. For example, consider a gamble in which the benefits outweighs the costs. OpAL* predicts that decision makers should more consistently opt to take this gamble when it is presented in the context of a rich environment. Indeed, this is precisely what was found by economist researchers, who also considered such patterns to be indicative of efficient coding (***Frydman and Jin, 2021***).

In this study, participants were presented with a series of trials where they selected between a gamble with a varying magnitude X with 50% probability and a certain option with varying magnitude C. The task featured two conditions, which we refer to as Rich and Lean. The range (minimum and maximum) of X's and C's were equated across the two conditions, but high magnitude Xs and Cs were more frequent in the Rich environment, whereas low magnitude Xs and Cs were more frequent in the Lean environment. The distribution of C was set to 0.5*X so that the expected value of the risky lottery and certain option were on average equated. Critically, there were a few carefully selected "common trials" that repeated the exact same high payoff gambles (with identical X and C) across blocks. The authors reported that participants were more likely to gamble on common trials in Rich environments than Lean environments. This is in line with their economic efficient-coding model, which predicts subjects allocate more resources to accurately perceive higher payoffs in the Rich condition where higher payoffs are more frequent (and therefore gamble more on common trials which are high payoff).

To simulate this dataset with OpAL*, we assumed that the critic state value would reflect the statistics of the environment. We first set the baseline expectation to reflect the expected value of a uniform prior over the gamble magnitudes and certain magnitudes in the experiment, which serves as a prior for environment richness. ρ was modulated by the learned average gamble offer in the environment relative to this baseline. [7] As in our earlier risky choice simulations, gambles

[7] This reference-dependent modulation is analogous to our learning experiments, in which the implicit baseline used a mean reward probability of 50%, and where environments with higher estimated reward probabilities were considered "rich" and those below 50% were considered "lean".

One could more generally apply the terms "rich" and "lean" to any values which deviate from a determined baseline, where $\bar{R}$ represents the estimated richness of the current environment and B represents the mean of an uninformative prior over the expected outcomes, $\rho(t) \propto \bar{R(t)} - B$. $\rho \geq 0$ would be considered "rich"; $\rho < 0$ would be considered "lean". Indeed, previous work has suggested that of a single environment may be encoded by tonic levels of dopamine, inducing changes in vigor of actions (***Niv et al., 2007***), but does not model changes in the choices themselves as we do here. A similar approach is used in average reward reinforcement learning. Rather than maximizing the total cumulative reward, average reward RL additionally optimizes the average reward per timestep. Reward prediction errors are therefore computed relative to the long-term average

were encoded relative to the certain option and G/N values were explicitly set according to the instructed gamble, omitting the need again for annealing and actor learning rate while preserving the core dynamics of the full OpAL*. As found empirically and in the authors' efficient coding model ***Frydman and Jin*** (***2021***), OpAL* predicts increased gambling on common trials in the Rich block relative to the Lean block. According to OpAL*, this result reflects adaptively modulated DA levels in the Rich environment which emphasized the benefits of the gamble during decision-making. As will be discussed below, OpAL*'s amplification of one striatal subpopulation over another itself can be considered a form of efficient coding, offering a direct mechanistic explanation for recent findings in economic theory. Finally, note that such findings could not be captured by an alternative model in which risky choice is driven by surprise or novelty. Note that for both rich and lean blocks, common trials had larger than usual magnitudes of payoffs. While these payoffs deviated from expectation to a larger degree in the lean block, this should produce a larger RPE (and presumably phasic dopamine signal). Given that increased DA in traditional RL models promotes exploitation (***Humphries et al., 2012***), this account (like the RSRL model above) would predict the opposite pattern than that seen empirically, in this case driving more risky choices in the lean block.

Discussion

Taken together, our simulations provide a normative account for opponency within the basal ganglia and its modulation by DA. In particular, we suggest that nonlinear Hebbian mechanisms give rise to convexity in the learned D1 and D2 actor weights at different ends of the reward spectrum, which can be differentially leveraged to adapt decision making. To do so, OpAL* alters its dopaminergic state as a function of environmetal richness, so as to best discern between the costs or benefits of available options. Conjecturing that such a mechanism is most profitable when the reward statistics of the environment are unknown, we posited and found that the online adaptation robustly outperforms traditional RL and alternative BG models across environment types when sampling across a wide range of plausible parameters. These advantages grow monotonically with the complexity of the environment (number of alternative actions to choose from). Morever, the unity of all three key features of OpAL* (opponency, three-factor Hebbian nonlinearity, and dynamic DA modulation) offered particularly unique advantages in sparse reward environments, mitigating against a particularly pernicious explore exploit dilemma that arises in such environments. Finally, we showed how such a mechanism can adapt risky decision making according to environmental richness, capturing the impact of DA manipulations and individual differences thereof.

This paper intersects with theoretical (***Niv et al., 2007***) and empirical work (***Hamid et al., 2015***; ***Mohebi et al., 2019***) investigating how changes in dopaminergic state reflecting reward expectations impact motivation and vigor. However, this body of literature does not consider how increases or decreases of dopamine affect the decision itself, only its latency or speed. Instead, OpAL/OpAL* can capture both shifts in vigor and cost-benefit choice as seen empirically with drug manipulations across species (***Cousins et al., 1996***; ***Salamone et al., 2005***; ***Treadway et al., 2012***; ***Westbrook et al., 2020***) and more precise optogenetic manipulations of DA and activity of D1 and D2 MSNs (***Doi et al., 2020***; ***Bolkan et al., 2022***; ***Zalocusky et al., 2016***; ***Tai et al., 2012***; ***Yartsev et al., 2018***). Notably, OpAL* suggests that in sparse reward environments, it is *adaptive* to lower dopaminergic levels and not merely avoiding action altogether (as in classical notions of the direct indirect pathways). Rather, lower dopamine helps to choose actions that minimize cost (by discriminating between D2 MSN populations). In physical effort decision tasks, DA depletion does not simply induce more noise or reduced effort overall, but selectively promotes actions that minimize effort when the benefits of exerting effort are relatively low (***Cousins et al., 1996***). For example, while a healthy rat will choose to climb a barrier to obtain four pellets instead of selecting two pellets that do not require physical effort, a dopamine depleted animal will opt for the two pellet-option.

reward per time step ($\bar{r}$), resulting in $\delta(t) = r(t) - \bar{r} - V(t)$. ρ as operationalized in OpAL* resembles a prediction error at the task/environment level, though may additionally be influenced by trial-by-trial prediction errors when trials are sufficiently distinct as in the interleaved gambles in ***Rutledge et al.*** (***2015***).

However, in the absence of the two pellet option, both healthy and dopamine depleted animals will select to climb the barrier to collect their reward. While OpAL* naturally accounts for such findings, other models often suggest that lowered DA levels would simply produce more randomness and imprecision, as captured by a reduced softmax gain (***FitzGerald et al., 2015***; ***Cinotti et al., 2019***; ***Eisenegger et al., 2014***; ***Lee et al., 2015***). Importantly, empirical evidence for this reduced gain account in low DA situations focused exclusively on reward rich situations (i.e., available options were likely to be rewarding); in these cases OpAL* also predicts more noise. But as noted above, low dopaminergic states may not always be maladaptive. Indeed, they may be useful in environments with sparse rewards, allowing an agent to adaptively navigate exploration and exploitation and to avoid the most costly options.

The work described here builds off a preliminary suggestion in ***Collins and Frank*** (***2014***) that opponency in OpAL confers advantages over standard RL models across rich environments and lean environments. In particular, when parameters were optimized for each model, the optimal parameters for standard RL diverged across environments, whereas OpAL could maximize rewards across environments with a single set of parameters; biological agents have indeed demonstrated similar learning speeds between lean and rich environments, demonstrating such cross-environment flexibility (***Hamid et al., 2015***). However, this previous work applied to a balanced OpAL model and did not consider how an agent might adaptively modulate dopaminergic state to differentially weigh costs vs benefits of alternative decisions. In this paper, we showed that such advantages are robust across a wide range of parameters (not just optimal ones), that they are amplified in OpAL*, and that such advantages grow with the complexity of the environment (number of alternative actions). Importantly, such benefits of OpAL* capitalize on the nonlinear and opponency convexity induced by Hebbian plasticity within D1 and D2 pathways (Figure 6a).

These findings contrast with other theoretical models of striatal opponency which omit the Hebbian term but leverage alternate nonlinearities so that, under certain parameter settings, D1 and D2 weights converge to the veridical benefits and costs of an action (***Möller and Bogacz, 2019***). However, for this convergence to occur requires assumptions about some knowledge of the reward distributions of the environment in advance. Our approach here is to consider how a model might optimize performance across variable environments with no fore knowledge; as such, OpAL* showed robust advantages over these alternative formulations. Such advantages *depended* on the nonlinear Hebbian mechanism. While the Hebbian term was originally motivated by the biology of three-factor plasticity as implemented in the neural network version, it is also needed to capture findings in which D2 MSNs become increasingly potentiated as a result of pathological DA depletion or DA blockade, leading to aberrant behavioral learning and progression of Parkinsonism (***Wiecki et al., 2009***; ***Beeler et al., 2012***). Ironically, it is this same Hebbian-induced nonlinearity that affords adaptive performance in OpAL* when DA is not depleted or manipulated exogenously.[8] Finally, this adaptive role for activity-dependent Hebbian plasticity beyond standard learning algorithms is complementary to recent observations that such mechanisms can be leveraged to improve beyond gradient descent in neural networks (***Scott and Frank, 2021***). While the computations are leveraged for different purposes (roughly, choice vs. credit assignment) and in different architectures, both findings accord with the notion that mechanisms typically thought to merely approximate adaptive functions inspired by artificial intelligence may in fact confer benefits for biological agents.

Lastly, while many studies have documented *that* DA manipulations affect risky and effort based decision making across species, our results offer a normative explanation for such findings. In this perspective, the brain treats increases or decreases in dopamine as signaling presence in a richer or leaner state. Changes in behavior reflect an adaption to this perceived, artificial environmental change. Hence, a dopamine depleted animal (or increased activity of D2 MSNs in ***Zalocusky et al.*** (***2016***)) would focus on costs of actions, whereas dopamine increases would increase attractiveness

[8] While (***Möller and Bogacz, 2019***) identified situations in which this mechanism can produce pathological behavior even without DA depletion, OpAL* rescues this behavior via normalization and annealing (see Appendix).

of risky actions (***Rutledge et al., 2015***). We reasoned that the well known impact of exogenous DA modulation on risky decision making (***St Onge and Floresco, 2009***; ***Zalocusky et al., 2016***; ***Rutledge et al., 2015***) may be a byproduct of this endogenous adaptive mechanism, showing that OpAL* can be used to modulate appropriately when it is worth taking a risk (Figure 8). We then demonstrated how behavioral effects of D2-receptor activity and manipulation (***Zalocusky et al., 2016***) reflect unique predictions of OpAL*, including outcome-dependent risk-avoidance paired with increase of D2-activity following a loss (Figure 9 A-D). In conjunction, optogenetic stimulation of D2-expressing neurons induced decrease in risky choice in risk-seeking rodents in line with OpAL* predictions (Figure 9 E-F). Furthermore, we showed that OpAL* can be used to capture changes in risk taking by dopamine-enhancing medication in healthy human participants (Figure 10, A-B). Our simulations highlighted how individual changes in risk preference may emerge from OpAL*'s adaptive mechanism. While some studies have shown that in unique circumstances increased dopamine may result in preference for a low-risk but low-reward option (***Mikhael and Gershman, 2021***; ***St. Onge et al., 2010***), these results rely on sequential effects but nonetheless they may be explainable by OpAL*'s sensitivity to environmental reward statistics. Furthermore, we focused on adaptive decision-making on the time scale of a single task in this paper, and it is plausible that such an adaptive mechanism may account for larger individual differences across longer time horizons. For example, increased risk-taking has been well documented in adolescents and some evidence suggests that dopaminergic levels may peak during adolescents, attributing to this trend (see (***Wahlstrom et al., 2010a***) for a full review). Speculatively, this may itself be an adaptive mechanism, where higher DA may allow more emphasis on potential benefits of risky but developmentally beneficial actions, such as exploring outside of parent's home to find a mate.

OpAL*'s separation and selective amplification of G and N actors also is reminiscent of efficient coding principles in sensory processing, which theorizes that neurons maximize information capacity by minimizing redundancy in neural representations (***Barlow, 2012***; ***Laughlin, 1981a***; ***Chalk et al., 2018***). Efficient coding also suggests that resources should be reallocated according to features in an environment which occur more frequently (***Simoncelli and Olshausen, 2001***). In the case of OpAL*, positive prediction errors are more abundant than negative in reward rich environments and the G actor strengthens disproportionately as this asymmetry grows. Conversely, negative prediction errors are more frequent in reward lean environments and the N actor specializes in this asymmetry. Changes in dopaminergic state, which modifies the contribution of G and N actors, therefore reallocates decision making resources according to the relative frequency of positive and negative prediction errors in the environment. Recent behavioral work has applied an efficient coding framework to risky choice paradigms, showing participants are riskier in environments which have an increased frequency of large gamble payoffs (***Frydman and Jin, 2021***). Our model provides a mechanistic account of such findings that generalizes to broader behavioral implications. Moreover, while the authors did not test this pattern, OpAL* predicts that if common trials were administered to include unfavorable gambles (gambles whose expected values are less than a certain option), people would more reliably select the certain outcome in the lean environment.

Limitations and future directions

A limitation of the DA modulation mechanism is that its performance advantages depend on relatively accurate estimates of environmental richness. Indeed, performance can suffer with incorrect estimation of the environment richness (Appendix, Figure 11). Thus it is essential in OpAL* that DA modulation is dynamic across trials so as to reflect sufficient reward history before modulating opponency. As such, while we systematically characterized the advantage of dynamic DA modulation in OpAL* over the balanced OpAL model ($\rho = 0$) across environments, this advantage should hold over any OpAL model with a fixed asymmetry (see Figure 6a). For robust advantages, the critic estimation of environmental richness must be relatively confident before modulating DA. In the simulations presented, we utilized a Bayesian critic to explicitly track such uncertainty, and only

increasing or decreasing DA when the estimate was sufficiently confident. Interestingly, this mechanism provides an intermediate strategy between directed and random exploration (***Wilson et al., 2014***), but at the level of actor (rather than action) selection. In OpAL*, such a strategy amounts to random exploration across both actors until the critic uncertainty is sufficiently reduced, at which point OpAL* exploits the actor most specialized to the task. Future directions will investigate how this strategy may itself be adapted as a function of the environment statistics and may offer potential predictions for understanding individual differences and/or clinical conditions. For example, given inappropriate dopaminergic state is most detrimental to sparse reward environments, an agent which prioritizes avoidance of costs such as those prevalent in sparse reward environments (such as in OCD or in early life stress) may benefit from more caution before changing dopaminergic state (i.e., have a higher threshold for DA modulation and exploiting knowledge) or take longer to integrate information to increase precision of estimates (i.e., lower learning rate).

There are several future directions to this work. For example, while OpAL* optimizes a single DA signal toward the actor most specialized to rich or lean environments, recent work also suggests that DA signals are not uniform across striatum (***Hamid et al., 2021***). Indeed, this work showed that DA signals can be tailored to striatal subregions specialized for a given task, keeping with a "mixture of experts" model to support credit assignment. Future work should thus consider how the DA signals can be simultaneously adapted to the benefits and costs of alternative actions within subregions that are most suited to govern behavior. Moreover, while we addressed the impact of complexity within the action space, an alternative notion of complexity and sparsity yet to be explored is the length of sequential actions needed to achieve reward. Increasing the distance from initial choice to reward, a problem faced by modern deep RL algorithms (***Hare, 2019***), may also benefit from integrating OpAL*-like opponency and choice modulation into larger architectures. Finally, while our work focuses on asymmetries afforded in the choice function, DA manipulations can also induce asymmetries in learning rates from positive and negative RPEs (***Frank et al., 2007a***; ***Niv et al., 2012a***; ***Collins and Frank, 2014***), which can, under some circumstances, be dissociated from choice effects (***Collins and Frank, 2014***). However, it is certainly possible that asymmetries in learning rates can also be optimized as a function of the environment. Indeed, larger learning rates for positive than negative RPEs are beneficial in lean environments (and vice-versa), by amplifying the less frequent signal (***Cazé and van der Meer, 2013***). Such effects are not mutually exclusive with those described here, but note that they do not address the issue highlighted above with respect to exploration exploitation dilemmas that arise in lean environments, and do not capture the various findings (reviewed above) in which DA manipulations affect performance and choice in absence of outcomes.

Materials and Methods

Parameter grid search

We ran a grid sweep over a parameter space with $\alpha_a \in [.1, 1]$ with step size of .1 and $\beta \in [1, 5]$ with step size of 0.5. To equate the model complexity, the annealing parameter ($T = 10$), the strength of modulation ($k = 20$), and the confidence needed before modulation ($\phi = 1.0$) were fixed to the specified values across models. These were determined by coarser grid searches of the parameter space for reasonable performance of control models. For each parameter combination, we matched the starting random seed for three models – OpAL*, OpAL* with $\rho = 0$, and OpAL* with no three-factor hebbian term (No Hebb). For each parameter setting for each model type, we calculated the average softmax probability of selecting the best option (80% in rich environments or 30% in lean environments) across 5,000 simulations for 500 trials. We then took the area under the curve (AUC) of this averaged learning curve for different time horizons (50, 100, 250, 500 trials) and took the difference between the AUCs of OpAL* and OpAL* with $\rho = 0$ or OpAL* No Hebb of matched parameters. We conducted a one sample t-test on these differences, where a difference of zero was the null hypothesis.

We conducted the same set of analyses with the learning curves for the actual rewards received and received mirror results. We therefore only report the analysis according to the probability of selecting an action which is a finer grain measure of average performance.

Moller and Bogacz 2019 model

The Moller and Bogacz model (***Möller and Bogacz, 2019***) offers another computational account of how benefits and costs may be encoded in the D1/D2 striatal sub-populations. First note that this model defines benefits and costs as the *absolute magnitude* of positive and negative outcome for each action. In contrast, benefits and costs as represented in OpAL/OpAL* are relative metrics that relate to the proportion of positive and negative prediction errors in an environment (accordingly, for gamble simulations, an outcome of 0 is encoded as a cost relative to the sure thing, similar to other models of reference dependence). Second, both OpAL and Moller and Bogacz's model have nonlinearities in the learning rule (otherwise, as seen in our balanced OpAL model, the two pathways are redundant). However, rather than use Hebbian plasticity, Moller and Bogacz transform the prediction error itself (such that the impact of negative prediction errors is smaller in the G actor, and vice versa, parametrized by ϵ), and impose a weak decay (λ), as expressed below.

$$\Delta G(s,a) = \alpha f_{\epsilon}(\delta) - \lambda G(s,a) \tag{28}$$

$$\Delta N(s,a) = \alpha f_{-\epsilon}(\delta) - \lambda N(s,a) \tag{29}$$

$$f_{\epsilon} = \begin{cases} \delta & \text{for} \quad \delta > 0 \\ \epsilon\delta & \text{for} \quad \delta < 0 \end{cases} \tag{30}$$

Under certain assumptions about reward distributions and associated parameters, this learning rule allows the G and N weights to converge to the expected payoffs and costs of alternative actions. However, as noted above, we are interested here in the general case where reward statistics are not known in advance, and as such we simulated behavior from this model across a range of parameters, as we did for the other agents, but we also optimized its parameters (see below).

To select between actions, we used a softmax policy. While ***Möller and Bogacz*** (***2019***) explicitly do not use a softmax function in their simulations, they did so only because they were simulating behaviors in which an action may not be selected at all (i.e., they did not subject their agent to choose between different actions). In contrast, for all of our experiments, our agents must select an action each trial. We therefore generate a choice as follows using the softmax function by using the value of the action, $V(a)$.

$$V(a) = \frac{1}{2}(G - N) \tag{31}$$

$$p(a) = \frac{e^{\beta V(a)}}{\sum_{i \in A} e^{\beta V(i)}} \tag{32}$$

Optimized Models

For each model and for a given set of parameters, the average softmax probability of selecting the best option for 100 trials was calculated over 1000 simulations in each environment. The mean performance in rich and lean were then also averaged. Parameters which maximized this final average were found using SciPy's *differential_evolution* routine. For plotting, 1000 random seeds were generated and preserved across all models to start each simulation to minimize model differences due to noise.

For the standard Q-learner, the two free parameters – the learning rate (α) and the softmax temperature (β) – were optimized. Learning rates were bounded between 0 and 1. Softmax temperatures could range between 1 and 50. Optimized parameters were found to be $\alpha = 0.16, \beta = 46.86$.

For each version of OpAL* optimized – OpAL*, OpAL* with $\rho = 0$, OpAL* with no hebbian term – only the learning rate (α) and the softmax temperature (β) were optimized. As in the grid search analyses, the annealing parameter ($T = 10$), the strength of modulation ($k = 20$), and the confidence needed before modulation ($\phi = 1.0$) were fixed to equate model complexity and to speed convergence of the optimization routine. The softmax temperature was also bounded in the optimization routine between 1 and 5 to ensure model stability. After optimized values were found, small deviations in T, k and ϕ were run to ensure results did not rely on the selection of these exact parameters. Optimized parameters were as follows: OpAL*, $\alpha = .84, \beta = 2.43$. OpAL* with $\rho = 0$, $\alpha = 0.96, \beta = 4.13$. OpAL* with no Hebbian term, $\alpha = .88, \beta = 2.90$.

The Moller et al. model was optimized over all four free parameters. See Moller and Bogacz 2019 model for an overview of the model. In order for the model to converge to expected payoffs and costs, the decay parameter (λ) must be close to 0 and smaller than the learning rate and the nonlinearity parameter (ϵ) must be approximately 1. The authors offer a practical way to determine these constraints by first defining $c_q \approx 1$ and $c_s = 1$, where c_q and c_s derive from the equilibrium equations for the mean spread s and the mean q of rewards in the environment if G and N are to converge to expected payoffs and costs. By first selecting c_q and c_s close to one and selecting a learning rate α, ϵ and λ can be calculated as follow:

$$\epsilon = \frac{1 - c_s(1/c_q - 1)}{1 + c_s(1/c_q - 1)} \tag{33}$$

$$\lambda = \frac{\alpha(1 - \epsilon)}{2c_s} \tag{34}$$

In order to optimize Moller et al., c_q and c_s were bounded between 0.7 and 1 and α ranged between 0 and 1. The above equations were then used to calculate ϵ and λ during the optimization procedure. Like the standard Q-learner, the softmax temperature was bounded between 1 and 50. Optimized parameters: $\alpha = .07$, $\epsilon = .91$, $\lambda = .004$, $\beta = 30$ using $c_q = .95$ and $c_s = .88$.

Acknowledgments

AJ was partly supported by NIMH training grant T32MH115895 (PI's: Frank, Badre, Moore). The project was also supported by NIMH R01 MH084840-08A1 and NIMH P50 MH119467-01. Computing hardware was supported by NIH Office of the Director grant S10OD025181.

References

Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma;. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5635655/.

Barlow HB. Possible Principles Underlying the Transformations of Sensory Messages. In: Rosenblith WA, editor. *Sensory Communication* The MIT Press; 2012.p. 216–234. http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262518420.001.0001/upso-9780262518420-chapter-13, doi: 10.7551/mitpress/9780262518420.003.0013.

Beeler J, Frank M, McDaid J, Alexander E, Turkson S, Sol Bernandez M, McGehee D, Zhuang X. A Role for Dopamine-Mediated Learning in the Pathophysiology and Treatment of Parkinson's Disease. Cell Reports. 2012 Dec; 2(6):1747–1761. https://linkinghub.elsevier.com/retrieve/pii/S2211124712004111, doi: 10.1016/j.celrep.2012.11.014.

Bolkan SS, Stone IR, Pinto L, Ashwood ZC, Iravedra Garcia JM, Herman AL, Singh P, Bandi A, Cox J, Zimmerman CA, Cho JR, Engelhard B, Koay SA, Pillow JW, Witten IB. Strong and opponent contributions of dorsomedial striatal pathways to behavior depends on cognitive demands and task strategy. Nature Neuroscience. 2022 Mar; 25:345–357. https://doi.org/10.1038/s41593-022-01021-9, doi: 10.1038/s41593-022-01021-9.

Cazé RD, van der Meer MAA. Adaptive properties of differential learning rates for positive and negative outcomes. Biological Cybernetics. 2013 Dec; 107(6):711–719. http://link.springer.com/10.1007/s00422-013-0571-5, doi: 10.1007/s00422-013-0571-5.

875 **Chalk M**, Marre O, Tkačik G. Toward a unified theory of efficient, predictive, and sparse coding. Proceedings
876 of the National Academy of Sciences. 2018 Jan; 115(1):186–191. https://www.pnas.org/content/115/1/186,
877 doi: 10.1073/pnas.1711114115, publisher: National Academy of Sciences Section: Biological Sciences.

878 **Cinotti F**, Fresno V, Aklil N, Coutureau E, Girard B, Marchand AR, Khamassi M. Dopamine blockade impairs
879 the exploration-exploitation trade-off in rats. Scientific Reports. 2019 May; 9:6770. https://www.ncbi.nlm.nih.
880 gov/pmc/articles/PMC6494917/, doi: 10.1038/s41598-019-43245-z.

881 **Collins AGE**, Frank MJ. Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on
882 reinforcement learning and choice incentive. Psychological Review. 2014; 121(3):337–366. http://doi.apa.
883 org/getdoi.cfm?doi=10.1037/a0037015, doi: 10.1037/a0037015.

884 **Cousins MS**, Atherton A, Turner L, Salamone JD. Nucleus accumbens dopamine depletions alter relative re-
885 sponse allocation in a T-maze cost/benefit task. Behavioural Brain Research. 1996 Jan; 74(1-2):189–197.
886 https://linkinghub.elsevier.com/retrieve/pii/0166432895001514, doi: 10.1016/0166-4328(95)00151-4.

887 **Daw ND**, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Networks. 2002
888 Jun; 15(4-6):603–616. https://linkinghub.elsevier.com/retrieve/pii/S0893608002000527, doi: 10.1016/S0893-
889 6080(02)00052-7.

890 **Dayan P**, Balleine BW. Reward, Motivation, and Reinforcement Learning. Neuron. 2002 Oct; 36(2):285–298.
891 https://linkinghub.elsevier.com/retrieve/pii/S0896627302009637, doi: 10.1016/S0896-6273(02)00963-7.

892 **Doi T**, Fan Y, Gold JI, Ding L. The caudate nucleus contributes causally to decisions that balance reward
893 and uncertain visual information. eLife. 2020 Jun; 9:e56694. https://elifesciences.org/articles/56694, doi:
894 10.7554/eLife.56694.

895 **Dunovan K**, Verstynen T. Believer-Skeptic Meets Actor-Critic: Rethinking the Role of Basal Ganglia Pathways
896 during Decision-Making and Reinforcement Learning. Frontiers in Neuroscience. 2016 Mar; 10. http://journal.
897 frontiersin.org/Article/10.3389/fnins.2016.00106/abstract, doi: 10.3389/fnins.2016.00106.

898 **Eisenegger C**, Naef M, Linssen A, Clark L, Gandamaneni PK, Müller U, Robbins TW. Role of Dopamine D2
899 Receptors in Human Reinforcement Learning. Neuropsychopharmacology. 2014 Sep; 39(10):2366–2375.
900 https://www.nature.com/articles/npp201484, doi: 10.1038/npp.2014.84, bandiera_abtest: a Cg_type: Nature
901 Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term:
902 Decision Subject_term_id: decision.

903 **FitzGerald THB**, Dolan RJ, Friston K. Dopamine, reward learning, and active inference. Frontiers in Com-
904 putational Neuroscience. 2015; 9:136. https://www.frontiersin.org/article/10.3389/fncom.2015.00136, doi:
905 10.3389/fncom.2015.00136.

906 **Frank MJ**, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals
907 multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of
908 Sciences. 2007 Oct; 104(41):16311–16316. http://www.pnas.org/cgi/doi/10.1073/pnas.0706111104, doi:
909 10.1073/pnas.0706111104.

910 **Frank MJ**, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals
911 multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of
912 Sciences. 2007 Oct; 104(41):16311–16316. http://www.pnas.org/cgi/doi/10.1073/pnas.0706111104, doi:
913 10.1073/pnas.0706111104.

914 **Frank MJ**. Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational Account of Cogni-
915 tive Deficits in Medicated and Nonmedicated Parkinsonism. Journal of Cognitive Neuroscience. 2005 Jan;
916 17(1):51–72. https://direct.mit.edu/jocn/article/17/1/51-72/3948, doi: 10.1162/0898929052880093.

917 **Frank MJ**, Claus ED. Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, deci-
918 sion making, and reversal. Psychological Review. 2006; doi: 10.1037/0033-295X.113.2.300.

919 **Frank MJ**, Seeberger LC, O'Reilly RC. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinson-
920 ism. Science. 2004 Dec; 306(5703):1940–1943. https://www.science.org/doi/10.1126/science.1102941, doi:
921 10.1126/science.1102941.

922 **Franklin NT**, Frank MJ. A cholinergic feedback circuit to regulate striatal population uncertainty and op-
923 timize reinforcement learning. eLife. 2015 Dec; 4:e12029. https://elifesciences.org/articles/12029, doi:
924 10.7554/eLife.12029.

925 **Frydman C**, Jin LJ. Efficient Coding and Risky Choice. . 2021; p. 74.

926 **Frémaux N**, Gerstner W. Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor
927 Learning Rules. Frontiers in Neural Circuits. 2016 Jan; 9. http://journal.frontiersin.org/Article/10.3389/fncir.
928 2015.00085/abstract, doi: 10.3389/fncir.2015.00085.

929 **Gerfen CR**. The neostriatal mosaic: multiple levels of compartmental organization. . 1992; 15(4):7.

930 **Gurney KN**, Humphries MD, Redgrave P. A New Framework for Cortico-Striatal Plasticity: Behavioural Theory
931 Meets In Vitro Data at the Reinforcement-Action Interface. PLoS Biology. 2015 Jan; 13(1):e1002034. https:
932 //dx.plos.org/10.1371/journal.pbio.1002034, doi: 10.1371/journal.pbio.1002034.

933 **Hamid AA**, Frank MJ, Moore CI. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assign-
934 ment. Cell. 2021 May; 184(10):2733–2749.e16. https://linkinghub.elsevier.com/retrieve/pii/S0092867421003779,
935 doi: 10.1016/j.cell.2021.03.046.

936 **Hamid AA**, Pettibone JR, Mabrouk OS, Hetrick VL, Schmidt R, Vander Weele CM, Kennedy RT, Aragona BJ, Berke
937 JD. Mesolimbic dopamine signals the value of work. Nature Neuroscience. 2015; doi: 10.1038/nn.4173.

938 **Hare J**. Dealing with Sparse Rewards in Reinforcement Learning. arXiv:191009281 [cs, stat]. 2019 Nov; http:
939 //arxiv.org/abs/1910.09281, arXiv: 1910.09281.

940 **Harun R**, Hare KM, Brough EM, Munoz MJ, Grassi CM, Torres GE, Grace AA, Wagner AK. Fast-scan
941 cyclic voltammetry demonstrates that L-DOPA produces dose-dependent, regionally selective bi-
942 modal effects on striatal dopamine kinetics in vivo. Journal of Neurochemistry. 2016; 136(6):1270–
943 1283. https://onlinelibrary.wiley.com/doi/abs/10.1111/jnc.13444, doi: 10.1111/jnc.13444, _eprint:
944 https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnc.13444.

945 **Humphries M**, Khamassi M, Gurney K. Dopaminergic control of the exploration-exploitation trade-off via the
946 basal ganglia. Frontiers in Neuroscience. 2012; 6. https://www.frontiersin.org/article/10.3389/fnins.2012.00009.

947 **Iino Y**, Sawada T, Yamaguchi K, Tajiri M, Ishii S, Kasai H, Yagishita S. Dopamine D2 receptors in discrimination
948 learning and spine enlargement. Nature. 2020 Mar; 579(7800):555–560. http://www.nature.com/articles/
949 s41586-020-2115-1, doi: 10.1038/s41586-020-2115-1.

950 **Kahneman D**, Tversky A. Prospect Theory: An Analysis of Decision under Risk. Econometrica. 1979 Mar;
951 47(2):263. https://www.jstor.org/stable/1914185?origin=crossref, doi: 10.2307/1914185.

952 **Laughlin S**. A Simple Coding Procedure Enhances a Neuron's Information Capacity. Zeitschrift
953 für Naturforschung C. 1981 Oct; 36(9-10):910–912. https://www.degruyter.com/document/doi/10.1515/
954 znc-1981-9-1040/html, doi: 10.1515/znc-1981-9-1040.

955 **Laughlin S**. A Simple Coding Procedure Enhances a Neuron's Information Capacity. Zeitschrift
956 für Naturforschung C. 1981 Oct; 36(9-10):910–912. https://www.degruyter.com/document/doi/10.1515/
957 znc-1981-9-1040/html, doi: 10.1515/znc-1981-9-1040.

958 **Lee E**, Seo M, Monte OD, Averbeck BB. Injection of a Dopamine Type 2 Receptor Antagonist into the Dorsal Stria-
959 tum Disrupts Choices Driven by Previous Outcomes, But Not Perceptual Inference. Journal of Neuroscience.
960 2015 Apr; 35(16):6298–6306. https://www.jneurosci.org/content/35/16/6298, doi: 10.1523/JNEUROSCI.4561-
961 14.2015, publisher: Society for Neuroscience Section: Articles.

962 **Lloyd K**, Dayan P. Tamping Ramping: Algorithmic, Implementational, and Computational Explanations of Pha-
963 sic Dopamine Signals in the Accumbens. PLOS Computational Biology. 2015 Dec; 11(12):e1004622. https:
964 //journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004622, doi: 10.1371/journal.pcbi.1004622,
965 publisher: Public Library of Science.

966 **Maia TV**, Frank MJ. An Integrative Perspective on the Role of Dopamine in Schizophrenia. Biologi-
967 cal Psychiatry. 2017 Jan; 81(1):52–66. https://linkinghub.elsevier.com/retrieve/pii/S0006322316324258, doi:
968 10.1016/j.biopsych.2016.05.021.

969 **Mikhael JG**, Bogacz R. Learning Reward Uncertainty in the Basal Ganglia. PLoS Computational Biology. 2016;
970 doi: 10.1371/journal.pcbi.1005062.

971 **Mikhael JG**, Gershman SJ. Impulsivity and risk-seeking as Bayesian inference under dopaminergic con-
972 trol. Neuropsychopharmacology. 2021 Aug; https://www.nature.com/articles/s41386-021-01125-z, doi:
973 10.1038/s41386-021-01125-z.

Mohebi A, Pettibone JR, Hamid AA, Wong JMT, Vinson LT, Patriarchi T, Tian L, Kennedy RT, Berke JD. Dissociable dopamine dynamics for learning and motivation. Nature. 2019 Jun; 570(7759):65–70. http://www.nature.com/articles/s41586-019-1235-y, doi: 10.1038/s41586-019-1235-y.

Montague P, Dayan P, Sejnowski T. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of Neuroscience. 1996 Mar; 16(5):1936–1947. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.16-05-01936.1996, doi: 10.1523/JNEUROSCI.16-05-01936.1996.

Möller M, Bogacz R. Learning the payoffs and costs of actions. PLOS Computational Biology. 2019 Feb; 15(2):e1006285. https://dx.plos.org/10.1371/journal.pcbi.1006285, doi: 10.1371/journal.pcbi.1006285.

Niv Y, Edlund JA, Dayan P, O'Doherty JP. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. Journal of Neuroscience. 2012 Jan; 32(2):551–562. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.5498-10.2012, doi: 10.1523/JNEUROSCI.5498-10.2012.

Niv Y, Edlund JA, Dayan P, O'Doherty JP. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. Journal of Neuroscience. 2012 Jan; 32(2):551–562. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.5498-10.2012, doi: 10.1523/JNEUROSCI.5498-10.2012.

Niv Y. Reinforcement learning in the brain. Journal of Mathematical Psychology. 2009 Jun; 53(3):139–154. https://linkinghub.elsevier.com/retrieve/pii/S0022249608001181, doi: 10.1016/j.jmp.2008.12.005.

Niv Y, Daw ND, Joel D, Dayan P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology. 2007 Mar; 191(3):507–520. http://link.springer.com/10.1007/s00213-006-0502-4, doi: 10.1007/s00213-006-0502-4.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 2006 Aug; 442(7106):1042–1045. http://www.nature.com/articles/nature05051, doi: 10.1038/nature05051.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 2006 Aug; 442(7106):1042–1045. http://www.nature.com/articles/nature05051, doi: 10.1038/nature05051.

Qi L, Thomas E, White SH, Smith SK, Lee CA, Wilson LR, Sombers LA. Unmasking the Effects of L-DOPA on Rapid Dopamine Signaling with an Improved Approach for Nafion Coating Carbon-Fiber Microelectrodes. Analytical Chemistry. 2016 Aug; 88(16):8129–8136. https://pubs.acs.org/doi/10.1021/acs.analchem.6b01871, doi: 10.1021/acs.analchem.6b01871.

Ratcliff R, Frank MJ. Reinforcement-Based Decision Making in Corticostriatal Circuits: Mutual Constraints by Neurocomputational and Diffusion Models. Neural Computation. 2012 May; 24(5):1186–1229. https://direct.mit.edu/neco/article/24/5/1186-1229/7757, doi: 10.1162/NECO_a_00270.

Reynolds JNJ, Wickens JR. Dopamine-dependent plasticity of corticostriatal synapses. Neural Networks. 2002 Jun; 15(4-6):507–521. https://linkinghub.elsevier.com/retrieve/pii/S089360800200045X, doi: 10.1016/S0893-6080(02)00045-X.

Rutledge RB, Skandali N, Dayan P, Dolan RJ. Dopaminergic Modulation of Decision Making and Subjective Well-Being. Journal of Neuroscience. 2015 Jul; 35(27):9811–9822. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0702-15.2015, doi: 10.1523/JNEUROSCI.0702-15.2015.

Salamone J, Correa M, Mingote S, Weber S. Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. Current Opinion in Pharmacology. 2005 Feb; 5(1):34–41. https://linkinghub.elsevier.com/retrieve/pii/S1471489204002000, doi: 10.1016/j.coph.2004.09.004.

Salamone JD, Correa M, Nunes EJ, Randall PA, Pardo M. The Behavioral Pharmacology of Effort-related Choice Behavior: Dopamine, Adenosine and Beyond. Journal of the Experimental Analysis of Behavior. 2012 Jan; 97(1):125–146. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3266736/, doi: 10.1901/jeab.2012.97-125.

Salamone JD, Correa M, Yang JH, Rotolo R, Presby R. Dopamine, Effort-Based Choice, and Behavioral Economics: Basic and Translational Research. Frontiers in Behavioral Neuroscience. 2018; 12:52. https://www.frontiersin.org/article/10.3389/fnbeh.2018.00052, doi: 10.3389/fnbeh.2018.00052.

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; doi: 10.1126/science.275.5306.1593.

Scott DN, Frank MJ. Beyond gradients: Noise correlations control Hebbian plasticity to shape credit assignment. bioRxiv. 2021; .

Simoncelli EP, Olshausen BA. Natural Image Statistics and Neural Representation. Annual Review of Neuroscience. 2001 Mar; 24(1):1193–1216. https://www.annualreviews.org/doi/10.1146/annurev.neuro.24.1.1193, doi: 10.1146/annurev.neuro.24.1.1193.

St Onge JR, Chiu YC, Floresco SB. Differential effects of dopaminergic manipulations on risky choice. Psychopharmacology. 2010 Aug; 211(2):209–221. https://doi.org/10.1007/s00213-010-1883-y, doi: 10.1007/s00213-010-1883-y.

St Onge JR, Floresco SB. Dopaminergic Modulation of Risk-Based Decision Making. Neuropsychopharmacology. 2009 Feb; 34(3):681–697. http://www.nature.com/articles/npp2008121, doi: 10.1038/npp.2008.121.

Sutton RS, Barto AG. Reinforcement Learning. Second ed. Cambridge, MA: MIT Press; 2018. http://incompleteideas.net/book/RLbook2020.pdf.

Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nature Neuroscience. 2012 Sep; 15(9):1281–1289. doi: 10.1038/nn.3188.

Tobler PN, Fiorillo CD, Schultz W. Adaptive coding of reward value by dopamine neurons. Science (New York, NY). 2005 Mar; 307(5715):1642–1645. doi: 10.1126/science.1105370.

Treadway MT, Buckholtz JW, Cowan RL, Woodward ND, Li R, Ansari MS, Baldwin RM, Schwartzman AN, Kessler RM, Zald DH. Dopaminergic Mechanisms of Individual Differences in Human Effort-Based Decision-Making. Journal of Neuroscience. 2012 May; 32(18):6170–6176. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.6459-11.2012, doi: 10.1523/JNEUROSCI.6459-11.2012.

Voon V, Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ, Hallett M. Mechanisms Underlying Dopamine-Mediated Reward Bias in Compulsive Behaviors. Neuron. 2010 Jan; 65(1):135–142. https://linkinghub.elsevier.com/retrieve/pii/S0896627309010459, doi: 10.1016/j.neuron.2009.12.027.

Wahlstrom D, Collins P, White T, Luciana M. Developmental changes in dopamine neurotransmission in adolescence: Behavioral implications and issues in assessment. Brain and Cognition. 2010 Feb; 72(1):146–159. https://linkinghub.elsevier.com/retrieve/pii/S027826260900205X, doi: 10.1016/j.bandc.2009.10.013.

Wahlstrom D, Collins P, White T, Luciana M. Developmental changes in dopamine neurotransmission in adolescence: Behavioral implications and issues in assessment. Brain and Cognition. 2010 Feb; 72(1):146–159. https://linkinghub.elsevier.com/retrieve/pii/S027826260900205X, doi: 10.1016/j.bandc.2009.10.013.

Westbrook A, van den Bosch R, Määttä JI, Hofmans L, Papadopetraki D, Cools R, Frank MJ. Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. Science. 2020; doi: 10.1126/science.aaz5891.

Wiecki TV, Riedinger K, von Ameln-Mayerhofer A, Schmidt WJ, Frank MJ. A neurocomputational account of catalepsy sensitization induced by D2 receptor blockade in rats: context dependency, extinction, and renewal. Psychopharmacology. 2009 Jun; 204(2):265–277. http://link.springer.com/10.1007/s00213-008-1457-4, doi: 10.1007/s00213-008-1457-4.

Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General. 2014; 143(6):2074–2081. http://doi.apa.org/getdoi.cfm?doi=10.1037/a0038199, doi: 10.1037/a0038199.

Yartsev MM, Hanks TD, Yoon AM, Brody CD. Causal contribution and dynamical encoding in the striatum during evidence accumulation. eLife. 2018 Aug; 7:e34929. https://doi.org/10.7554/eLife.34929, doi: 10.7554/eLife.34929, publisher: eLife Sciences Publications, Ltd.

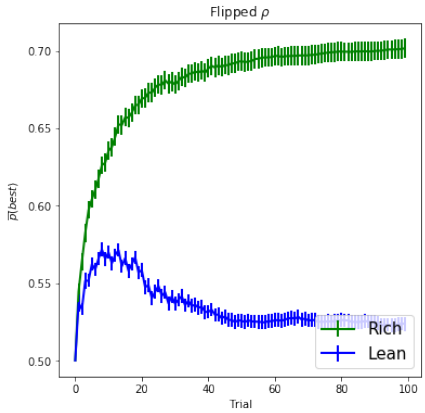
Yttri EA, Dudman JT. Opponent and bidirectional control of movement velocity in the basal ganglia. Nature. 2016; doi: 10.1038/nature17639.

Zalocusky KA, Ramakrishnan C, Lerner TN, Davidson TJ, Knutson B, Deisseroth K. Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making. Nature. 2016; doi: 10.1038/nature17400.

Appendix

Incorrect modulation impairs performance

As noted in the main text, it is important that the critic estimate of environmental richness is reasonably accurate (on the correct side of 0.5) for OpAL* to confer advantages. Indeed, pathological behavior arises if DA states are altered in opposing direction to environmental richness. In Appendix Figure 11 we see the effect of flipping the sign of OpAL*'s calculation of dopaminergic state (Equation 17). For this demonstration, if the critic of OpAL* estimated that it was in a rich environment (positive value of rho, high dopaminergic state), it would emphasize the N instead of G actor (as if it were in a lean environment). We see that the lean environment shows high sensitivity to incorrect modulation. The rich environment shows greater robustness but nonetheless has decreased performance in comparison to the standard simulations. This result confirms that the direction of modulation in OpAL* is important, and moreover that it is particularly important to have lower DA in lean environments.



Appendix 0 Figure 11. Effects of dopaminergic states which inaccurately reflect environmental richness. Parameters for OpAL* as outlined in the section Optimized Models were used for these simulations.

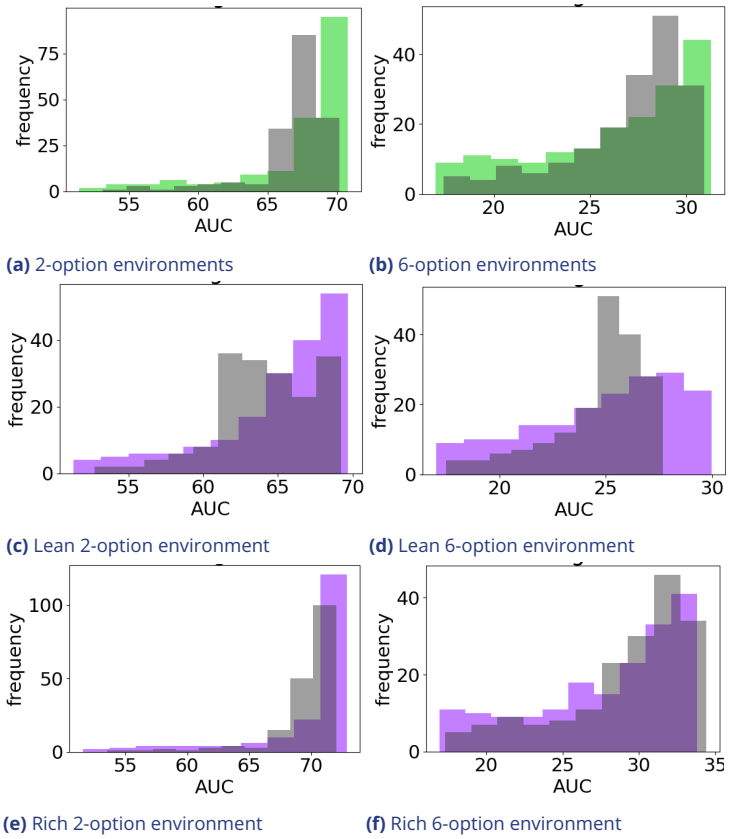
Comparison to softmax temperature modulation

As noted in the main text, OpAL* confers larger benefits in lean environments, in part by mitigating against an exploration/exploitation dilemma. In particular, during early learning, OpAL* relies on both actors equally and thereby distributes its policy more randomly, but after it estimates the richness of the environment, it exploits the more specialized actor. To evaluate whether similar benefits could be mimicked by simply increasing softmax gain over trials (transitioning from exploration to exploitation), we considered an OpAL* variant which symmetrically increased the softmax temperature according equally across the G and the N actor. As the richness (or leanness) of the environment grew, the agent would progressively exploit both actors equally, using the same Bayesian critic as in OpAL*.

$$\beta_g = \beta \max(0, 1 + |\rho_t|) \tag{35}$$

$$\beta_n = \beta \max(0, 1 + |\rho_t|) \tag{36}$$

Given the difference in exploration-exploitation demands across rich and lean environments, we compared the average AUCs of OpAL* and Beta-modulation (B-Mod). Overall we found that OpAL* exhibited improved maximal cross-environment robustness and specifically improved maximal performance in the lean environment. Thus, global changes in explore-exploit the softmax temperature alone are insufficient to capture the full performance benefit in lean environments induced by dopaminergic modulation in OpAL*, which capitalizes on specialized learned representations across actors.



Appendix 0 Figure 12. Comparison of OpAL* to dynamic modulation of softmax temperature (bmod). Figure shows average AUCs of models for fixed parameter in both lean and rich environments for varying complexity.Top: AUCs averaged across both rich and lean environments for a given parameter. Green - OpAL*, Grey - BMod. Middle/Bottom: AUC histograms for different environments and varying complexity levels. Purple - OpAL*, Grey - BMod.

Nonlinear dynamics forego veridicality for flexibility

Addressing *Möller and Bogacz* (*2019*)

We incorporated normalization and weight decay for the actors to address weaknesses of the original OpAL model raised by ***Möller and Bogacz*** (***2019***). The (valid) critique outlined by ***Möller and Bogacz*** (***2019***) is that its three-factor hebbian update, in carefully constructed situations, gives rise to unstable actor dynamics. They demonstrated that when OpAL is sequentially presented with a reward of 2 followed by a cost of -1, the dynamics of G and N rapidly converge to 0 (Figure 13, left). As described in their text (Equations 39-41), stable oscillations in reward prediction errors cause G and N values to converge towards zero. This is indeed a characteristic of the OpAL model, especially once the critic begins to converge.[9].

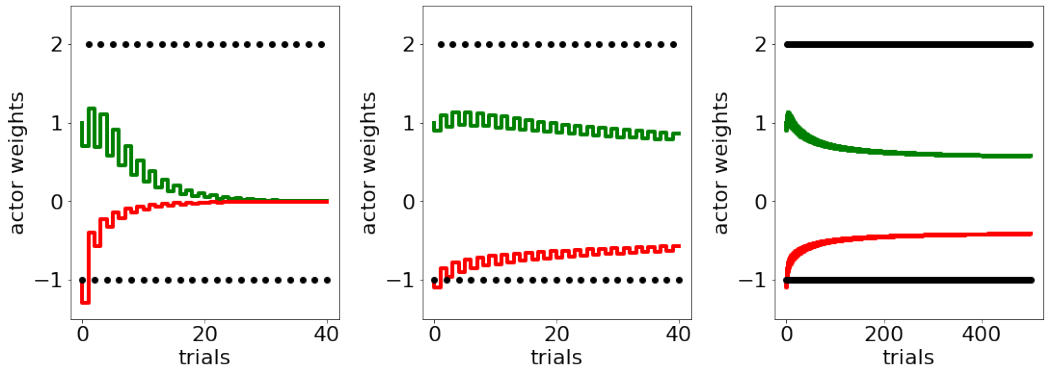
The rapid decay evident in Figure 13, left, was constructed to highlight a particularly pernicious example of this issue. The following simulations suggest that the introduction of larger reward magnitudes, rather than the oscillating PEs, have driven such instability. Larger reward magnitudes yield larger reward prediction error signals, that in turn yield larger G/N values as evident by Equations 4 and 5, which, through the Hebbian positive feedback cycle, further increase learning rate. One simple correction is to simply rescale and shrink the magnitudes by some constant ($0 < c < 1$); this slows decay in this example (simulations not shown).

We introduced two modifications in OpAL* to address these concerns. First, prediction errors used to update G and N actors (Equations 21 and 22) are normalized by the range of known reward magnitudes in the environment (Equation 24). Importantly, OpAL* is not provided any reward statistics beyond the range of reward feedback, and in theory this value could be adjusted as the agent learns, reflecting how dopamine neurons rapidly adapt to the range of reward values in the environment ***Tobler et al.*** (***2005***).

Figure 13, center, shows the effect of normalization for the example in question. We see that the rapid decay is substantially decreased, and simulating into a farther time horizon of 100 trials shows a trend toward, but not final convergence at, zero (Figure 13, right). (Note OpAL* behaves well for several hundred trials in the experiments we simulated in this paper). While there remains a general decay over time, as previously stated, the behavior is reminiscent of advantage learning curves, which have the positive feature that such decay can encourage the agent to explore after many trials in the event the world has changed. Furthermore, it is plausible that other learning mechanisms, such as more habitual stimulus-response learning, also contribute to choice after many learning trials (***Frank and Claus, 2006***). Thus striatal weight decay, which has been documented empirically (***Yttri and Dudman, 2016***), may not be detrimental for procedural performance. Normalizing, therefore, addresses the valid concerns of ***Möller and Bogacz*** (***2019***) while still preserving core OpAL dynamics, which allow it to capture a range of biological phenomenon as well as hypotheses for advantages of dopaminergic states presented in this paper.

Secondly, to address the original issue raised by ***Möller and Bogacz*** (***2019***) that OpAL weights decay with oscillating prediction errors, we introduced annealing of the actor learning rate. This is a common addition to reinforcement learning algorithms where the learning rate is large in early stages of learning to avoid local minimums and slowly decreases with time to protect values in later stages of learning from rapid updating. (To allow for change points in reward statistics, other mechanisms capturing the effects of cholinergic interneurons have been shown to be useful in BG networks and OpAL variants, ***Franklin and Frank*** (***2015***)). Figure 13, right, shows that while actor weights still decrease with the addition of annealing, they no longer converge to zero and lose all prior learning as demonstrated in ***Möller and Bogacz*** (***2019***).

[9] Arguably, this decay could be akin to an advantage-learning action value curve, such that once the critic begins to converge, the "advantage" of the option (difference between the action value and the average value of the environment) decreases overtime (***Dayan and Balleine, 2002***). In neural network versions of our BG model, striatal action selection is only required for early learning; once a policy is repeated sufficiently, the cortex can directly select an action in a stimulus-response fashion ***Frank and Claus*** (***2006***)



Appendix 0 Figure 13. $\alpha_c = .3$, $\alpha_a = .3$. Middle, Left figure: $T = 50$, normalization = 3